Advanced Machine Learning

Fall 2021

Prof. Joachim M. Buhmann

Final Exam

February 4th, 2022

First and last name: _____

Student ID number: _____

Signature: _____

General Remarks

- Please check that you have all 44 pages of this exam.
- You can acquire a maximum of 100 points.
- The exam lasts 180 minutes.
- Advice: Do not spend too much time on a single question. You do not need to secure 100 points to achieve the top grade.
- Remove all material which is not permitted by the examination regulations from your desk.
- Write your answers directly on the exam sheets. If you need more space, make sure you put your student ID number on top of each supplementary sheet.
- Immediately inform an assistant in case that you are not able to take the exam under regular conditions. Delayed complaints are not accepted.
- Attempts to cheat/defraud lead to immediate notification to the rector's office with a possible exclusion from the examination and it might entail judicial consequences.
- Use a **black** or a **blue** pen to answer the questions. Pencils or red/green colored pens are not allowed.
- Provide only one solution to each exercise. Invalid solutions have to be clearly and unambiguously cancelled.
- **Grading of true/false questions**: You will receive 1 point per correct answer, -1 point per incorrect answer, and 0 points for no answer with a minimum of 0 points per question.
- **Grading of multiple choice questions**: You receive 1 point per correct answer and 0 points per incorrect answer or unanswered question.

	Topic	Points	Points achieved	Checked
1	Density estimation	12		
2	Regression	9		
3	Bias-variance tradeoff	7		
4	Linear methods	8		
5	Bayesian information criterion	8		
6	Convex optimization	8		
7	SVMs	8		
8	Ensembles	8		
9	NP-Bayes	8		
10	PAC-learning	12		
11	Model selection	12		
Total		100		

Question 1: Density Estimation: Frequentist Linear Regression (12 pts)

Let $Y \in \mathbb{R}$ be a random output variable and $\mathbf{x} \in \mathbb{R}^d$ be a *fixed* vector of features. Assume the following regression model:

$$Y = \boldsymbol{\beta}^\top \mathbf{x} + \epsilon, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^d$ is a fixed vector of regression parameters and $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

- Derive the distribution of $Y \mid \mathbf{x}, \boldsymbol{\beta}$. Write down its mean and variance.

2 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

$Y \mid \mathbf{x}, \boldsymbol{\beta} \sim \mathcal{N}\left(\boldsymbol{\beta}^\top \mathbf{x}, \sigma^2\right)$. Since $Y = \boldsymbol{\beta}^\top \mathbf{x} + \epsilon$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$, Y is normally-distributed; $\mathbb{E}[Y] = \mathbb{E}\left[\boldsymbol{\beta}^\top \mathbf{x} + \epsilon\right] = \boldsymbol{\beta}^\top \mathbf{x} + \mathbb{E}[\epsilon] = \boldsymbol{\beta}^\top \mathbf{x}$; and $\mathbb{V}[Y] = \mathbb{V}\left[\boldsymbol{\beta}^\top \mathbf{x} + \epsilon\right] = 0 + \mathbb{V}[\epsilon] = \sigma^2$.

- Assume given a dataset $\mathcal{Z} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, with $y_i \sim \mathbf{P}(Y \mid \mathbf{x}_i, \boldsymbol{\beta})$. Write down the log-likelihood function $\log \mathbf{P}(\mathcal{Z} \mid \boldsymbol{\beta})$. You may write it up to the terms constant in $\boldsymbol{\beta}$. You may assume that $\{\mathbf{x}_i\}_{i=1}^n$ are fixed.

3 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

$\mathbf{P}\left(\mathcal{Z} \mid \boldsymbol{\beta}\right) = \prod_{i=1}^{n} p\left(y_i \mid \mathbf{x}_i,\, \boldsymbol{\beta}\right)$.

$\log \mathbf{P}\left(\mathcal{Z} \mid \boldsymbol{\beta}\right) = \log\left(\prod_{i=1}^{n} p\left(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}\right)\right) = \sum_{i=1}^{n} \log\left\{ \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{1}{2}\left(\frac{y_i - \boldsymbol{\beta}^\top \mathbf{x}_i}{\sigma}\right)^2} \right\} = \text{const}_1 +$

$\sum_{i=1}^{n} -\frac{1}{2\sigma^2}\left(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i\right)^2 = \text{const}_1 - \text{const}_2 \sum_{i=1}^{n}\left(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i\right)^2$.

- Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the design matrix with rows given by $\mathbf{x}_i$ from $\mathcal{Z}$ and let $\mathbf{y} = (y_1, ..., y_n)^\top$. Demonstrate that the MLE for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}^n_{ML} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$. You may assume $\mathbf{X}^\top \mathbf{X}$ to be invertible.

3 pts ☐

$\hat{\boldsymbol{\beta}}^n_{ML} = \arg\max_{\boldsymbol{\beta}} \mathbf{P}\left(\mathcal{Z} \,|\, \boldsymbol{\beta}\right) = \arg\max_{\boldsymbol{\beta}} \left\{ -\sum_{i=1}^n \left(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i\right)^2 \right\} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n \left(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i\right)^2.$
$\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n \left(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i\right)^2 = \arg\min_{\boldsymbol{\beta}} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 =: \arg\min_{\boldsymbol{\beta}} \mathcal{L}_{\boldsymbol{\beta}}.$

$\frac{\partial \mathcal{L}_{\boldsymbol{\beta}}}{\partial \boldsymbol{\beta}} = 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}^\top \mathbf{y}$. For $\hat{\boldsymbol{\beta}}^n_{ML}$, $2\mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\beta}}^n_{ML} - 2\mathbf{X}^\top \mathbf{y} \stackrel{!}{=} 0$.
$\hat{\boldsymbol{\beta}}^n_{ML} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$.

- What is the distribution of $\hat{\boldsymbol{\beta}}_{ML}^n$? Derive the distribution's parameters and show that $\hat{\boldsymbol{\beta}}_{ML}^n$ is an unbiased estimator of $\boldsymbol{\beta}$.

4 pts ☐

Recall that $Y \,|\, \mathbf{x}, \boldsymbol{\beta} \sim \mathcal{N}\left(\boldsymbol{\beta}^\top \mathbf{x}, \sigma^2\right)$. Then $\hat{\boldsymbol{\beta}}_{ML}^n$ should follow a multivariate Gaus-

sian distribution with the mean $\mathbb{E}\left[\hat{\boldsymbol{\beta}}_{ML}^{n}\right] = \mathbb{E}\left[\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{y}\right] = \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbb{E}\left[\mathbf{y}\right] = \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbb{E}\left[\mathbf{X}\boldsymbol{\beta}+\epsilon\right] = \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\left(\mathbf{X}\boldsymbol{\beta}+\mathbb{E}\left[\epsilon\right]\right) = \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$ and covariance $\mathbb{V}\left[\hat{\boldsymbol{\beta}}_{ML}^{n}\right] = \mathbb{V}\left[\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{y}\right] = \mathbb{V}\left[\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\left(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}\right)\right] = 0+\mathbb{V}\left[\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\boldsymbol{\epsilon}\right] = \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{X}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbb{V}\left[\boldsymbol{\epsilon}\right] = \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbb{I}_{d}\sigma^{2} = \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\sigma^{2}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}_{d}\left(\mathbf{0},\, \mathbb{I}_{d}\sigma^{2}\right)$ is a vector with i.i.d. noise terms. Thus, $\hat{\boldsymbol{\beta}}_{ML}^{n} \sim \mathcal{N}_{d}\left(\boldsymbol{\beta},\, \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\sigma^{2}\right)$. $\hat{\boldsymbol{\beta}}_{ML}^{n}$ is an unbiased estimator $\boldsymbol{\beta}$, since $\mathbb{E}\left[\hat{\boldsymbol{\beta}}_{ML}^{n}\right] = \boldsymbol{\beta}$.

Grading scheme a

- -2 (-1) if (minor) conceptual mistake / incorrect derivation
- -0.5 if mixed scalar and matrix notation (i.e. gave $I_d\sigma$)

b

- -1 for not logging
- -2 if given general $p(y|x)$ form, unless expanded later
- -1 if given $N(y_i|x_i)$, unless expanded later
- -0.5 for minor errors in derivation (signs, no $\propto$)
- -1 if mix scalar and matrix notation

c

- -2 if they invert with x^T
- -1 for minor derivation mistakes (eg signs, transpose mistakes)

d

- $+2$ for $E[\beta]$ and $+1$ for just unbiased definition
- $+2$ for $Var[\beta]$. -1 if $Var[\beta]$ is derived incorrectly. -2 if larger mistakes in derivation (eg not understanding matrix properties)

Question 2: Regression (9 pts)

We continue with the same notation as in Question 1.

- Are the following claims true or false?

4 pts ☐

1. Consider the regression model $Y = f(\mathbf{x}, \boldsymbol{\beta}) + \epsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \beta_3 \log x_3 + \epsilon$ with i.i.d. error terms $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We can obtain an unbiased estimator of the coefficients $\boldsymbol{\beta}$ by fitting the linear least-squares regression on the appropriately transformed x_1, x_2, x_3.
   ☐ True ☐ False

   True.
2. For any two random variables X and Y $\arg\min_g \mathbb{E}[(Y - g(X))^2] = \mathbb{E}[Y \mid X]$.
   ☐ True ☐ False

   True.
3. Assume that $Y = \mathbf{x}^\top \boldsymbol{\beta} + \epsilon$. The least-squares estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ has the smallest variance among all estimators of $\boldsymbol{\beta}$ of the form $\mathbf{C}\mathbf{y}$, for some $\mathbf{C} \in \mathbb{R}^{d \times n}$
   ☐ True ☐ False

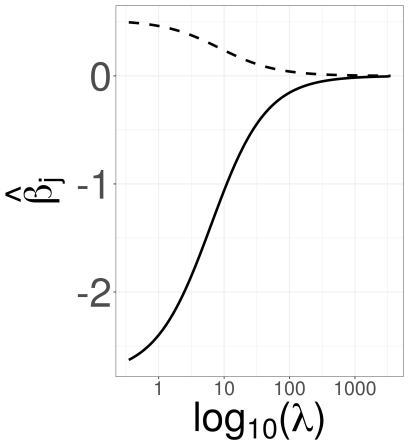
   False. Consider the ridge estimator.
4. For the least-squares estimator $\hat{\boldsymbol{\beta}}$ above to exist, no column of the design matrix $\mathbf{X}$ must be a linear combination of other columns.
   ☐ True ☐ False

   True.

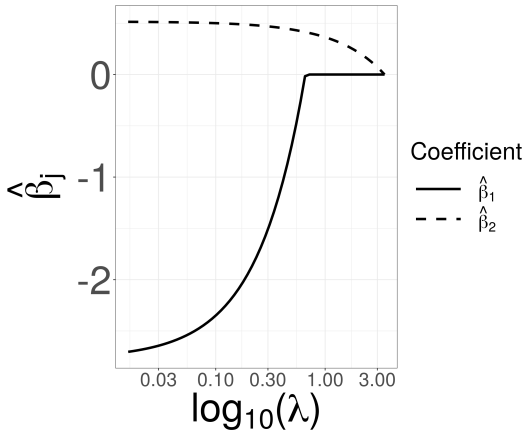
- Choose the correct answer.

5 pts ☐

1. Consider LASSO and ridge regression with two features. Which of the plots below shows the estimated regression coefficients $\hat{\boldsymbol{\beta}}$ across varying regularization parameter λ for the LASSO regression (one plot correponds to LASSO and another to ridge)? *Note: each curve corresponds to the values of a single estimated coefficient across a range of regularization parameter values.*



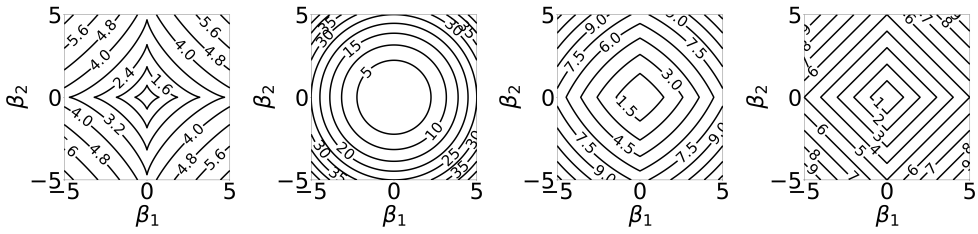
(a) ☐



(b) ☐

(b)

2. Consider the closed form solution for ridge regression $\hat{\boldsymbol{\beta}}^{\text{ridge}} = \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbb{I}_d\right)^{-1}\mathbf{X}^\top\mathbf{y}$. The estimator $\hat{\boldsymbol{\beta}}^{\text{ridge}}$ follows
   ☐ a normal distribution with the mean $\boldsymbol{\beta}$
   ☐ a normal distribution with the mean $\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbb{I}_d\right)^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}$
   ☐ an unknown distribution with the mean $\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbb{I}_d\right)^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}$
   A normal distribution with the mean $\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbb{I}_d\right)^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}$.
3. LASSO regression is equivalent to the MAP estimator in Bayesian linear regression with the prior on β_i given by
   ☐ a Laplace distribution with the location parameter 0 and scale parameter $\frac{2\sigma^2}{\lambda}$
   ☐ a normal distribution with the mean 0 and variance $\frac{\sigma^2}{\lambda}$
   ☐ a Cauchy distrubition with the mean 0 and variance $\frac{\sigma^2}{\lambda}$
   A Laplace distribution with the location parameter 0 and scale parameter $\frac{2\sigma^2}{\lambda}$.
4. Consider $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1\right\}$, where $\lambda_1, \lambda_2 \geq 0$ are regularization parameters. Which of the plots below depicts the level sets of the penalty term $\lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1$ for 2 features, $\lambda_2 = 0.1$, and $\lambda_1 = 0.9$?



(a) ☐ (b) ☐ (c) ☐ (d) ☐

(c)

5. An astrologist wants to predict the life expectancy of her clients. She considers computing the least-squares estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ for linear regression with an intercept term and four binary features coding whether a client was born in the corresponding season (spring, summer, autumn, and winter). Her dataset contains at least 5 clients. Which statement is true?
   ☐ The least-squares estimator exists, however, the ridge regression estimator would have a lower test MSE in this setting.
   ☐ The least-squares estimator exists, however, least-squares linear regression is not appropriate for datasets with categorically-valued features.
   ☐ The least-squares estimator does not exist, since the features are collinear.
   The least-squares estimator does not exist, since the features are collinear.

Question 3: Bias-variance tradeoff (7 pts)

Consider a classification problem where we want to predict labels $y \in \mathcal{Y}$ from features $x \in \mathcal{X}$ using a finite data set $D_n = \{(x_i, y_i)_{i=1}^n\}$. We use the term *estimator* to refer to any function $f : \mathcal{X} \to \mathcal{Y}$.

- Let $\hat{f}_{ERM}$ be the empirical risk minimizer over a certain hypothesis class. List three methods to reduce the variance of $\hat{f}_{ERM}$.

1 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Collecting more data, increasing/adding regularization, increasing model size well beyond interpolation (double descent), reducing model complexity, performing feature selection. Grading: +0.5 first answer, +0.25 second answer, +0.25 third answer. No negative points.

- Are the following claims true or false?

6 pts ☐

1. Reducing the bias of any estimator $f : \mathcal{X} \to \mathcal{Y}$ increases its variance.
   ☐ True ☐ False

   False.
2. Consider an SVM estimator that uses an RBF kernel $k(x, z) = \exp(-\gamma\|x - z\|^2)$. Decreasing the value of the coefficient γ leads to a lower variance.
   ☐ True ☐ False

   True.
3. If an estimator interpolates the training data (i.e. achieves 0 training error) then it has poor generalization (i.e. high test error).
   ☐ True ☐ False

   False.
4. Boosting and bagging reduce both the bias and the variance of the individual estimators.
   ☐ True ☐ False

   False.
5. Consider a finite-sample data set with noiseless samples. In this case, interpolating the data is sufficient for good generalization.
   ☐ True ☐ False

   False.
6. Increasing the sample size n reduces the variance of any estimator $f : \mathcal{X} \to \mathcal{Y}$.
   ☐ True ☐ False

False, e.g. the empirical count estimator for number of heads of a coin toss.

Question 4: Linear Methods (8 pts)

Assume given a dataset $\mathcal{Z} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$ where $\mathbf{x}_i \in \mathbb{R}^d$ are features and $y_i \in \{0, 1\}$ are labels, for $1 \leq i \leq n$. Let $\mathcal{C}_0 = \{i : \ y_i = 0, \ 1 \leq i \leq n\}$ and $\mathcal{C}_1 = \{i : \ y_i = 1, \ 1 \leq i \leq n\}$.

- Which of the following classification methods are **(a)** generative, **(b)** probabilistic discriminative, **(c)** discriminative? Put the correct letter ('a', 'b' or 'c') next to each method.

2 pts ☐

1. Perceptron _____
2. Logistic regression _____
3. A maximum likelihood approach modeling class-conditional densities and class priors _____
4. Fisher's linear discriminant with a threshold for classifying projected data points _____

1 – c, 2 – b, 3 – a, 4 – c.

- Assume given class-conditional densities for features $\mathbf{x}$ denoted by $p(\mathbf{x} \mid y = k)$ and class prior probabilities $p(y = k)$, for $k \in \{0, 1\}$. Derive an expression for the posterior probability $p(y = 0 \mid \mathbf{x})$ in terms of class-conditional densities and class prior probabilities.

1 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

By Bayes' theorem, $p(y = 0 \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid y=0)p(y=0)}{p(\mathbf{x})} = \frac{p(\mathbf{x} \mid y=0)p(y=0)}{p(\mathbf{x} \mid y=0)p(y=0)+p(\mathbf{x} \mid y=1)p(y=1)}$.

For the remainder of this question, we will focus on Fisher's linear discriminant. Recall that Fisher's linear discriminant is given by a weight vector $\mathbf{w} \in \mathbb{R}^d$ maximizing the criterion

$$J(\mathbf{w}) = \frac{\left(\mathbf{w}^\top \mathbf{m}_1 - \mathbf{w}^\top \mathbf{m}_0\right)^2}{\sum_{k \in \{0,1\}} \sum_{i \in \mathcal{C}_k} \left(\mathbf{w}^\top \mathbf{x}_i - \mathbf{w}^\top \mathbf{m}_k\right)^2}, \tag{2}$$

where $\mathbf{m}_k = \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \mathbf{x}_i$, for $k \in \{0, 1\}$. We will consider a classifier constructed by applying a threshold to the projected data points $\mathbf{w}^\top \mathbf{x}_i$.

- Explain the meaning of the numerator and denominator in the criterion $J(\mathbf{w})$ above.

2 pts

. . . . . .

. . . . . .

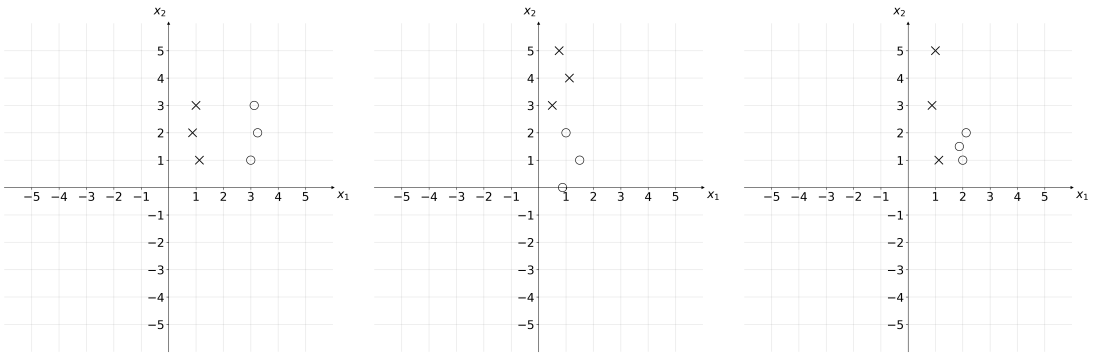
. . . . . .

. . . . . .

Numerator corresponds to the separability of the projected class centroids; denominator corresponds to the scatter of the projected data points within each class.

- Consider two alternative criteria $J^{(1)}(\mathbf{w}) = \left(\mathbf{w}^\top \mathbf{m}_1 - \mathbf{w}^\top \mathbf{m}_0\right)^2$ and $J^{(2)}(\mathbf{w}) = \frac{1}{\sum_{k\in\{0,1\}} \sum_{i\in\mathcal{C}_k} \left(\mathbf{w}^\top \mathbf{x}_i - \mathbf{w}^\top \mathbf{m}_k\right)^2}$. The plots below depict three different datasets. Symbols '$\times$' and '$\circ$' denote classes 0 and 1, respectively. For $\mathbf{w}$ maximizing $J^{(1)}(\mathbf{w})$, for which dataset would the projected data points $\mathbf{w}^\top \mathbf{x}_i$ not be linearly separable? For which dataset would the projected data points $\mathbf{w}^\top \mathbf{x}_i$ not be linearly separable when maximizing $J^{(2)}(\mathbf{w})$? Write down the letter of the corresponding dataset next to the appropriate criterion.

1 pts



(a) (b) (c)

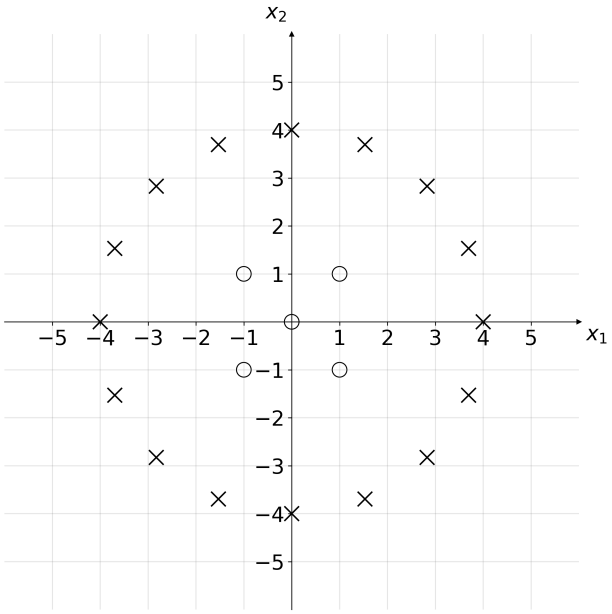
$J^{(1)}(\cdot)$ ____

$J^{(2)}(\cdot)$ ____

$J^{(1)}(\mathbf{w})$ fails for (c); $J^{(2)}(\mathbf{w})$ fails for (b).

- Consider another dataset plotted below. How many solutions are there maximizing Fisher's criterion $J(\mathbf{w})$? Justify your answer.

2 pts



There are infinitely many optimal projection lines. Any $\mathbf{w} \in \mathbb{R}^2$ maximizes $J(\mathbf{w})$. Observe that $J(\mathbf{w}) = 0$ for any $\mathbf{w} \in \mathbb{R}^2$, since $\mathbf{m}_0 = \mathbf{m}_1 = \mathbf{0}$ and the numerator of $J(\mathbf{w})$ is always equal-to 0.

Grading scheme

- $+0.5$ points for each correct answer. No penalty for the wrong one.
- Full 1 point for the correct formula. 0.5 points if demominator is only written as p(x).
- $+$ 1 points for correct answer for each meaning. It should not be exact as in the solution but the idea of what nomination and denominator mean should be correct.

- $+0.5$ points for each correct answer. No penalty for the wrong one.
- $+1$ points for correct answer (infinitely many solutions) and $+1$ point for the correct explanation. If the answer is wrong then 0 points.

Question 5: Bayesian information criterion (BIC) (8 pts)

All subquestions here are independent. You do not need to have solved the previous ones to solve the next.

Assume that a dataset $\mathbf{X}$ is given and that we want to estimate the parameter of a model $\{p(\cdot \mid \theta) \mid \theta \in \mathbb{R}^k\}$. In this question, we ask you to reproduce the derivation of the BIC in the lecture. The BIC is defined by the formula $-\log p(\mathbf{X} \mid \theta^{ML}) + \frac{k}{2} \log n$, where n is the number of examples in the dataset $\mathbf{X}$ and k is the number of entries in θ^{ML}, the maximum likelihood estimator.

We start from a Bayesian perspective. Write $p(\mathbf{X})$ in terms of the likelihood $p(\mathbf{X} \mid \theta)$ and a prior $p(\theta)$, here θ is a variable denoting a parameter.

1 pts ☐

. . . . . .

. . . . . .

$$p(\mathbf{X}) = \int p(\mathbf{X} \mid \theta) p(\theta) d\theta. \tag{3}$$

Grading: +1 points for correct solution and +0.25 points if sum is written instead of integral. 0 points otherwise.

Assume that we use a very flat prior. Demonstrate that

$$p(\mathbf{X}) \approx \text{const } n^{-k/2} p(\mathbf{X} \mid \theta^{ML}). \tag{4}$$

Hint: You may use the following facts without proof.

- $-I_n(\theta)$ is the Hessian of $\log p(\mathbf{X} \mid \theta)$ with respect to θ, where $I_n(\theta)$ is the Fisher information computed for $\mathbf{X}$ and θ. You may assume it is invertible.
- The normalization constant of $\mathcal{N}(\mu, \Sigma)$ is $|2\pi\Sigma|^{1/2}$.
- $\left|I(\theta^{ML})\right|$ is a constant.
- You may assume that second-order Taylor approximations are exact.
- It can be shown that the prior $p(\cdot)$ can be replaced with a constant over $\mathbb{R}^k$.

6 pts ☐

. . . . . .

. . . . . .

. . . . . .

The Taylor approximation is the following:

$$\log p(\mathbf{X} \mid \theta) \approx \log p(\mathbf{X} \mid \theta^{ML}) + \left(\theta - \theta^{ML}\right)^\top \nabla \log p(\mathbf{X} \mid \theta^{ML}) - \frac{1}{2}\left(\theta - \theta^{ML}\right)^\top I_n(\theta^{ML})\left(\theta - \theta^{ML}\right). \tag{5}$$

The linear term vanishes as $\nabla \log p(\mathbf{X} \mid \theta^{ML}) = 0$. Replacing this into Equation 3 and substituting $p(\theta)$ with a constant yields

$$\begin{aligned}
p(\mathbf{X}) &\approx \text{const} \int \exp\left(\log p(\mathbf{X} \mid \theta^{ML}) - \frac{1}{2}\left(\theta - \theta^{ML}\right)^\top I_n(\theta^{ML})\left(\theta - \theta^{ML}\right)\right) d\theta && (6)\\
&= \text{const}\; p(\mathbf{X} \mid \theta^{ML}) \int \exp\left(-\frac{1}{2}\left(\theta - \theta^{ML}\right)^\top I_n(\theta^{ML})\left(\theta - \theta^{ML}\right)\right) d\theta && (7)\\
&= \text{const}\; p(\mathbf{X} \mid \theta^{ML}) \left|2\pi I_n^{-1}(\theta^{ML})\right|^{1/2} && (8)\\
&= \text{const}\; p(\mathbf{X} \mid \theta^{ML}) \left|\frac{2\pi}{n} I^{-1}(\theta^{ML})\right|^{1/2} && (9)\\
&= \text{const}\; p(\mathbf{X} \mid \theta^{ML}) n^{-k/2}. && (10)\\
& && (11)
\end{aligned}$$

Grading: 6 points to fully correct answer, for partial answers:

- For those who followed the above steps:
  - 2 points for Taylor approximation (0.5 for the first two terms and 1.5 for the last)

- 1 points for gradient vanishing
- 1 points for the first $p(\mathbf{X})$ approximation $p(\mathbf{X}) \approx \int \exp ...$
- 1 points each for the last two steps of above solution.
- The students didnot necessarily follow the above steps. For those cases, I stayed loyal to above grading and distributed the points according to "midstep grading" above.

• Some additional notes:

- If there is an error that is not very important and didnot propagate, I cut little points, else I cut more. For instance, for a step that is 1 point worth, if there is a small error (seemed as simple as a typo or careless mistake) I cut 0.25, if it is propagated to another step or used in wrong context then I cut 0.5 etc. If it is fully propagated or if there is no explanation, I sometimes didnot give any points (for instance sometimes some steps were skipped and where n and k come from was unclear. As their existence in the final expression are given by the question itself, if there is no explanation sometimes they lost half to full points).
- Anybody who only wrote sum or integral marginalisation, I gave 0.25 points

Derivation of the BIC

Use the approximation above to show that

$$p(\mathbf{X}) \approx \text{const} \exp\left(-\text{BIC}\right). \tag{12}$$

1 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

$$\begin{aligned} p(\mathbf{X}) &\approx \text{const } p(\mathbf{X} \mid \theta^{ML}) n^{-k/2} && (13)\\ &= \text{const } \exp\left(\log p(\mathbf{X} \mid \theta^{ML}) + \log n^{-k/2}\right) && (14)\\ &= \text{const } \exp\left(\log p(\mathbf{X} \mid \theta^{ML}) - \frac{k}{2}\log n\right) && (15)\\ &= \text{const } \exp\left(-\text{BIC}\right) && (16)\\ & && (17) \end{aligned}$$

Grading: Anybody who put a single (and correct) step in between const $p(\mathbf{X} \mid \theta^{ML}) n^{-k/2}$ and $\exp\left(-\text{BIC}\right)$ or give a verbal explanation (implies, by defn. etc) got the full 1 points. If only

const $p(\mathbf{X} \mid \theta^{ML})n^{-k/2} = \exp(-\text{BIC})$ is written without replacing "BIC" with its expression got 0.5 points. If there is no additional step then just writing out what questions asks without putting BIC in, then they get 0 points. Finally, if a middle step is there but written wrongly (some confusions with log, exp etc.) they get 0.5 points.

Question 6: Convex optimization (8 pts)

Let f, g_i, h_j, for $i \leq k$ and $j \leq \ell$, be functions mapping $\mathbb{R}^p$ to $\mathbb{R}$. Demonstrate *the weak duality property of the primal and the dual:*

$$\begin{array}{lll} \max\limits_{\lambda \in \mathbb{R}^k, \alpha \in \mathbb{R}^\ell} \min\limits_{w \in \mathbb{R}^p} \mathcal{L}(w, \lambda, \alpha) & \leq & \min\limits_{w \in \mathbb{R}^p} \quad f(w) \\ s.t. \qquad \alpha_j \geq 0 \text{ for } j \leq \ell. & & s.t. \quad \begin{array}{ll} g_i(w) & = 0 \text{ for } i \leq k \text{ and} \\ h_j(w) & \leq 0 \text{ for } j \leq \ell. \end{array} \end{array} \tag{18}$$

Here, $\mathcal{L}$ is the Lagrangian of the right-hand side of the inequality, λ_i and α_j are the Lagrange multipliers associated to g_i and h_j, respectively. You may assume that all minima and maxima of these optimization problems exist.

For convenience, define $F = \{w \in \mathbb{R}^p \mid g_i(w) = 0, h_j(w) \leq 0, \text{ for } i \leq k \text{ and } j \leq \ell.\}$.

6 pts

1. The definition of Lagrangian (1pt):

$$\mathcal{L}(w, \lambda, \alpha) = f(w) + \sum_{i=0}^{k} \lambda_i g_i(w) + \sum_{j=0}^{l} \alpha_j h_j(w). \tag{19}$$

2. Now we consider arbitrary λ, α where $\alpha_j \geq 0, j \leq l$ (we can also write $\lambda \in \mathbb{R}^k, \alpha \in \mathbb{R}^l_+$). For $w \in F$ we thus have (1pt):

$$\lambda_i g_i(w) = 0, \ i \leq k, \tag{20}$$
$$\alpha_j h_j(w) \leq 0, \ j \leq l. \tag{21}$$

3. This means for $w \in F$, we can bound $\mathcal{L}$ by f (1pt):

$$\begin{aligned} \mathcal{L}(w, \lambda, \alpha) &= f(w) + 0 + \sum_{j=0}^{l} \alpha_j h_j(w) \\ &\leq f(w) + 0 + 0 = f(w). \end{aligned} \tag{22}$$

4. Since this holds for $\forall w \in F$, we can do the same for the minimum (1pt):

$$\min_{w \in F} \mathcal{L}(w, \lambda, \alpha) \leq \min_{w \in F} f(w), \ \forall \lambda \in \mathbb{R}^k, \alpha \in \mathbb{R}^l_+. \tag{23}$$

5. Then we extend the domain on the left-hand side to $w \in \mathbb{R}^p$, using the fact that extending the domain cannot increase the minimum (1pt):

$$\min_{w \in \mathbb{R}^p} \mathcal{L}(w, \lambda, \alpha) \leq \min_{w \in F} \mathcal{L}(w, \lambda, \alpha), \ \forall \lambda \in \mathbb{R}^k, \alpha \in \mathbb{R}^l_+. \tag{24}$$

Connecting the previous two inequalities we have

$$\min_{w \in \mathbb{R}^p} \mathcal{L}(w, \lambda, \alpha) \leq \min_{w \in F} f(w), \ \forall \lambda \in \mathbb{R}^k, \alpha \in \mathbb{R}^l_+. \tag{25}$$

6. Note that the right hand side does not depend on λ and α, so the inequality holds when taking the maximum with respect to λ and α (1pt):

$$\max_{\lambda \in \mathbb{R}^k, \alpha \in \mathbb{R}^l_+} \min_{w \in \mathbb{R}^p} \mathcal{L}(w, \lambda, \alpha) \leq \min_{w \in F} f(w), \tag{26}$$

which is the weak duality inequality.

Grading notes (Xianyao): I look for each of the steps, and if some steps are incorrectly skipped, the points are not given (a check mark means one point, a half-check mark is half a point). Students often get the different domains of w mixed up, thus skipping steps incorrectly. And for each case where the domain of w is incorrect, I deduct 0.5 points. That is, if they use $w \in \mathbb{R}^k$ from the beginning, there will be deductions at steps 2, 3, 4, 5 (in this case step 5 is usually omitted) and 6 should they not correct the domain. Often the domain is not clearly stated, and I try my best to salvage some points by finding the domain according to context.
On the other hand, students don't necessarily follow the solution's steps. In that case, I try to follow their proof and find corresponding steps with the solution. If the proof isn't entirely correct, I give points according to the steps I can find.

Finally, typos or other minor mistakes that affect the proof, like writing a $\leq$ in place of a $\geq$ inadvertently or missing a $\sum$ sign in the Lagrange definition, are also worth -0.5 points each. A particular case is, if a student misses both $\sum$ signs in step 1, the student will get 0 points for the step.

Briefly explain the importance of this inequality for training SVMs in infinitely dimensional spaces.

2 pts

. . . . . .

. . . . . .

For SVMs, the primal problem cannot be efficiently solved if we are working with infinite dimensions. However, the dual is always finitely-dimensional as long as the dataset is finite and can be solved using quadratic programming.

Grading note (Xianyao): The main thing I was looking for was a contrast between the primal and the dual, things like "infeasible" vs. "feasible", "intractable" vs. "tractable", "more efficient", "easier to solve", "harder to train". If a contrast is established with proper explanation (finite vs. infinite dimensions) I normally give 2 points, unless there are major factual errors. For example, I actually don't know if the primal is *intractable* or just *inefficient* so I see either acceptable, but if a student says "the dual is an *upper* bound of the primal" (it should be a lower bound), that is worth -0.5 points.
If they only mention something we *can* do with the dual (e.g. we can use the kernel trick or we can transform the problem into finite dimensions) or the primal without contrasting them, I often give 1 point, as *can* indicates a subjective choice, whose benefit for SVM training is unclear. However, if they state that these "allow" or "enable" us to train SVM, I accept it as half a contrast because these words indicate that we would otherwise not be able to do so, and I usually give 1.5 points or even 2, based on the other statements.
In other cases, I try to look for something about dimensionality, feasibility and computation efficiency, but do not give more than 1 point.

Question 7: Ensemble Methods (8 pts)

Let $\hat{f}^{\text{ens}}(\cdot) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{(b)}(\cdot)$, where $\hat{f}^{(b)}$, for $1 \leq b \leq B$, denotes the base model trained using some randomized algorithm $\mathcal{A}$ on a sample $\mathcal{Z} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\} \subseteq \mathbb{R}^d \times \mathbb{R}$. For any $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ and any two different $1 \leq b, \tilde{b} \leq B$, we assume the following:

- $\mathbb{E}_{\mathcal{Z},\mathcal{A}}\left[\hat{f}^{(b)}(\mathbf{x})\right] = y$, *i.e.* base models are unbiased,
- $\mathbb{V}_{\mathcal{Z},\mathcal{A}}\left[\hat{f}^{(b)}(\mathbf{x})\right] = \sigma^2$, and
- $\text{Cov}_{\mathcal{Z},\mathcal{A}}\left(\hat{f}^{(b)}(\mathbf{x}), \hat{f}^{(\tilde{b})}(\mathbf{x})\right) = \rho$.

For $1 \leq b \leq B$ and arbitrary $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$, show that for $\mathbb{E}_{\mathcal{Z},\mathcal{A}}\left[\left(\hat{f}^{\text{ens}}(\mathbf{x}) - y\right)^2\right] \leq \mathbb{E}_{\mathcal{Z},\mathcal{A}}\left[\left(\hat{f}^{(b)}(\mathbf{x}) - y\right)^2\right]$ to hold, it must be that $\rho \leq \sigma^2$. *Hint: you may use without proof the bias–variance decomposition of the MSE.*

8 pts

Master solution:
Please note that subscripts are omitted for simplicity of notation.
Recall that $\mathbb{E}\left[\left(\hat{f}(\mathbf{x}) - y\right)^2\right] = \left(\text{Bias}\left(\hat{f}\right)\right)^2 + \mathbb{V}\left[\hat{f}\right]$. As $\text{Bias}(\hat{f}(x)) = 0$, $E_{\mathcal{Z},\mathcal{A}}\left[\left(\hat{f}^{(b)}(\mathbf{x}) - y\right)^2\right] = \mathbb{V}\left[\hat{f}^{(b)}(\mathbf{x})\right] = \sigma^2$.
Since individual base models are unbiased, $\text{Bias}\left(\hat{f}^{\text{ens}}\right) = \text{Bias}\left(\frac{1}{B}\sum_{b=1}^{B}\hat{f}^{(b)}\right) = \frac{1}{B}\sum_{b=1}^{B}\text{Bias}\left(\hat{f}^{(b)}\right) = \text{Bias}\left(\hat{f}^{(b)}\right) = 0$, using linearity of expectation. Therefore, $\mathbb{E}\left[\left(\hat{f}^{\text{ens}}(\mathbf{x}) - y\right)^2\right] = \mathbb{V}\left[\hat{f}^{\text{ens}}(\mathbf{x})\right]$.
Observe that
$\mathbb{V}\left[\hat{f}^{\text{ens}}(\mathbf{x})\right] = \mathbb{V}\left[\frac{1}{B}\sum_{b=1}^{B}\hat{f}^{(b)}(\mathbf{x})\right] = \frac{1}{B^2}\left[B\mathbb{V}\left[\hat{f}^{(b)}(\mathbf{x})\right] + \sum_{j,k=1:\, j\neq k}^{B}\text{Cov}\left(\hat{f}^{(j)}(\mathbf{x}), \hat{f}^{(k)}(\mathbf{x})\right)\right]$
$= \frac{1}{B^2}\left[B\sigma^2 + B(B-1)\rho\right]$.
We want to show $\mathbb{E}\left[\left(\hat{f}^{\text{ens}}(\mathbf{x}) - y\right)^2\right] \leq \mathbb{E}\left[\left(\hat{f}^{(b)}(\mathbf{x}) - y\right)^2\right]$ and thus (plugging in the formulas of respective variances),
$\frac{1}{B^2}\left[B\sigma^2 + B(B-1)\rho\right] \leq \sigma^2$.
$\sigma^2 + (B-1)\rho \leq B\sigma^2$.
$(B-1)\rho \leq (B-1)\sigma^2$.
$\rho \leq \sigma^2$.

Points distribution:

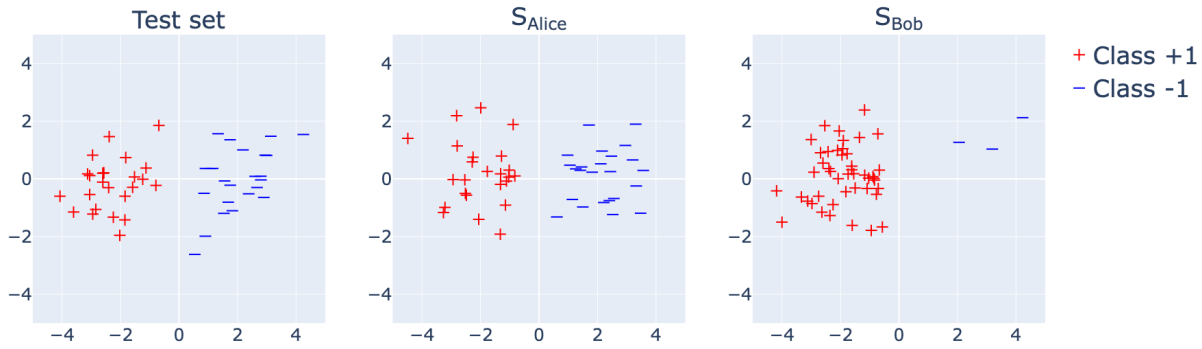
- MSE decomposition (bias^2 + Var) [1pt]
- Explicitly showing that ensemble is unbiased [1pt] ([0.5pt] if the fact stated without showing)

- Using the unbiasedness, i.e. MSE = Var [1pt]
- Rewriting $Var(\hat{f}^{ens}(x))$ in terms of $Var(\hat{f}^{b}(x))$ and covariance terms [1pt]
- Rewriting the above expression in terms of σ^2 and ρ [1pt]
- Simplifying the derived expression (make the logical flow to the conclusion) [1pt] ([0.5pt] if conclusion stated without any intermediate steps)
- Final conclusion ($\mathbb{E}_{\mathcal{Z},\mathcal{A}}\left[\left(\hat{f}^{\text{ens}}(\mathbf{x})-y\right)^2\right] \leq \mathbb{E}_{\mathcal{Z},\mathcal{A}}\left[\left(\hat{f}^{(b)}(\mathbf{x})-y\right)^2\right]$ iff $\rho \leq \sigma^2$) [1pt]
- Overall correctness of derivations [1pt]

Special cases and comments:

- All derivations correct except for the crucial part of rewriting $Var(\hat{f}^{ens}(x))$ (common mistake: $Var(\hat{f}^{ens}(x)) = \rho$) [4pt in total, as this is the core of the exercise and simplifies substantially all the following steps]
- All derivations correct but $\rho \leq \sigma^2$ used as assumption, instead of showing it, which was asked in the exercise [6 pt in total]
- $\mathbb{E}_{\mathcal{Z},\mathcal{A}}\left[\left(\hat{f}^{(b)}(\mathbf{x})-y\right)^2\right] = Var_{\mathcal{Z},\mathcal{A}}(\hat{f}^{b}(x))$ without any explanation [-0.5pt]
- Additionally provided incorrect information [-0.5pt]
- No explanation of why we look at $\frac{1}{B^2}\left[B\sigma^2 + B(B-1)\rho\right] \leq \sigma^2$ [-0.5pt]
- No punishment for omitting subscripts
- Both iff and one-way logical flow accepted, as the exercise formulation did not seem clear enough
- All derivations correct but no explanation and justification provided for the steps [-1pt or -3pt, depending on the extend]

Question 8: SVM (8 pts)



Consider the data sets in the figure above. Alice and Bob want to train an SVM using the data sets S_{Alice} and S_{Bob}, respectively, with $|S_{Alice}| = |S_{Bob}| = 50$. The test set and S_{Alice} are drawn from a distribution $\mathcal{D}_{Alice}$ with $P(Y = +1) = P(Y = -1) = \frac{1}{2}$, while S_{Bob} is drawn from distribution $\mathcal{D}_{Bob}$ with $\frac{P(Y=+1)}{P(Y=-1)} = 15$. Alice and Bob can choose between the following two types of SVMs:

$$\hat{f}_1 = \arg\min_w \frac{1}{2}\|w\|_2^2 + C_1 \sum_{i=1}^{n} \xi_i \qquad \text{(Soft-margin SVM)}$$
$$\text{s.t. } y_i w^T x_i \geq 1 - \xi_i \text{ and } \xi_i \geq 0, \forall i \in \{1, ..., n\}$$

$$\hat{f}_2 = \arg\min_w \frac{1}{2}\|w\|_2^2 + C_2 \left(C^{(+)} \sum_{\{i:y_i=+1\}} \xi_i + C^{(-)} \sum_{\{i:y_i=-1\}} \xi_i \right) \qquad \text{(Cost-sensitive SVM)}$$
$$\text{s.t. } y_i w^T x_i \geq 1 - \xi_i \text{ and } \xi_i \geq 0, \forall i \in \{1, ..., n\}$$

- Choose the correct answer.

5 pts ☐

1. How does increasing C_1 affect the number of support vectors of $\hat{f}_1(S_{Alice})$?
   ☐ the number of support vectors increases
   ☐ the number of support vectors decreases
   ☐ the number of support vectors is independent of C_1
   The number of support vectors decreases, as we recover hard margin SVM as C_1 grows to ∞.
2. Suppose that $C^{(+)}, C^{(-)} > 0$. How does increasing C_2 affect the number of support vectors of $\hat{f}_2(S_{Bob})$?
   ☐ the number of support vectors increases
   ☐ the number of support vectors decreases
   ☐ the number of support vectors is independent of C_2
   The number of support vectors decreases, as we recover hard margin SVM as C_2 grows to ∞.

3. Let $S^A \sim \mathcal{D}_{Alice}$ and $S^B \sim \mathcal{D}_{Bob}$ with $\left|S^A\right| = \left|S^B\right| = 50$. Which is true as $C_1 \to \infty$?
   ☐ $\hat{f}_1(S^A)$ and $\hat{f}_1(S^B)$ have the same bias, but $\hat{f}_1(S^A)$ has a higher variance.
   ☐ $\hat{f}_1(S^A)$ and $\hat{f}_1(S^B)$ have the same bias, but $\hat{f}_1(S^B)$ has a higher variance.
   ☐ $\hat{f}_1(S^B)$ has both higher bias and higher variance, compared to $\hat{f}_1(S^A)$.
   ☐ $\hat{f}_1(S^A)$ and $\hat{f}_1(S^B)$ have the same variance, but $\hat{f}_1(S^A)$ has a higher bias.
   ☐ $\hat{f}_1(S^A)$ and $\hat{f}_1(S^B)$ have the same variance, but $\hat{f}_1(S^B)$ has a higher bias.
   For $C_1 \to \infty$, we recover the hard margin SVM. The max margin estimator trained on S^B has higher variance, since it depends on fewer negative samples. The bias is also higher since the average estimator will be closer to the mean of the negative class conditional.
4. Assume that parameter values are set to $C_1 = C_2 = 1$. Identify the incorrect statement among the following:
   ☐ The test error of $\hat{f}_1(S_{Alice})$ is smaller than the test error of $\hat{f}_1(S_{Bob})$
   ☐ There exists a choice of $C^{(+)}$ and $C^{(-)}$ for which the test error of $\hat{f}_2(S_{Bob})$ is the same as that of $\hat{f}_1(S_{Alice})$
   ☐ Consider the $\hat{f}_2(S_{Alice})$ estimator obtained for the fixed values $C^{(+)} = 100$ and $C^{(-)} = 1$. The test error of $\hat{f}_2(S_{Alice})$ is smaller than the test error of $\hat{f}_1(S_{Alice})$.
   The last sentence is the false one.
5. How does increasing C_2 (think $C_2 >> C^{(+)}$ and $C_2 >> C^{(-)}$) affect the test error of $\hat{f}_2(S_{Bob})$?
   ☐ the test error increases
   ☐ the test error decreases
   ☐ the test error is independent of C_2
   The test error increases with C, as the balancing effect of $C^{(+)}$ and $C^{(-)}$ becomes dominated by the magnitude of C_2.

- Write the Lagrangian of the optimization problem for $\hat{f}_2(\mathcal{D}_{Bob})$.

3 pts ☐

$$\mathcal{L}(w, \alpha, \beta) = \frac{1}{2}\|w\|_2^2 + C_2 \left(C^{(+)} \sum_{\{i:y_i=+1\}} \xi_i + C^{(-)} \sum_{\{i:y_i=-1\}} \xi_i \right) + \sum_i \alpha_i (1 - \xi_i - y_i w^T x_i) - \sum_i \beta_i \xi_i$$

Grading scheme: Multiple choice:

- +1 for correct answer, 0 no answer

Lagrangian:

- -1 point for sign errors
- -2 points for missing constraints in the Lagrangian
- -2 points for wrong main objective

Question 9: Non-parametric Bayesian inference (8 pts)

Let $X = (X_1, \ldots, X_n)$ be a sample drawn from a non-parametric mixture of Gaussians. Let $Z = (Z_1, \ldots, Z_n)$ be random variables taking values in $\mathbb{N}$, according to a CRP. For $i \leq n$ and $k \in \mathbb{N}$, $\mathbf{P}(Z_i = k)$ denotes the probability that X_i should be assigned to cluster k. For $i \leq n$, denote by Z_{-i} the vector $(Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n)$. Define X_{-i} analogously.

In the lecture, we constructed a collapsed Gibbs sampler to estimate the posterior distribution. For this, we needed to compute $\mathbf{P}(Z_i = k \mid Z_{-i} = z, X = x, \theta)$, for some adequate z and $x = (x_1, \ldots, x_n)$. Here, θ are the priors' parameters, which we omit for convenience in the following.

Demonstrate that

$$\mathbf{P}\left(Z_i = k \mid Z_{-i} = z, X = x\right) \propto \mathbf{P}(Z_n = k \mid Z_{-n} = z) \times \tag{27}$$

$$\mathbf{P}(X_i = x_i \mid X_{-i} = x_{-i}, Z_i = k, Z_{-i} = z). \tag{28}$$

8 pts

The solution to the problem given in https://ml2.inf.ethz.ch/courses/aml/slides/aml21-lecture-13-npb.pdf – slide 46/53 and 47/53

Question 10: PAC learning (12 pts)

Consider a binary classification problem where the covariates are drawn from an arbitrary distribution $x \sim \mathcal{D}$ and the output variable is given by $y = c^*(x), \forall x \in \mathcal{X}$ where $c^* \in \mathcal{C}$. Given a finite data set $S = \{(x_i, y_i)\}_{i=1}^n$, we want to estimate $c^* \in \mathcal{C}$ with low error, using a hypothesis from $\mathcal{H}$.

- Are the following claims true or false?

  4 pts ☐

  1. The minimum number of samples required to PAC-learn concept class $\mathcal{C}$ using functions from hypothesis class $\mathcal{H}$ is determined by the VC dimension of set $\mathcal{C}$.
     ☐ True ☐ False
     False.
  2. Assume that $\mathcal{C} \neq \mathcal{H}$. It is necessary that $\mathcal{C} \subseteq \mathcal{H}$, in order for $\mathcal{C}$ to be PAC learnable from $\mathcal{H}$.
     ☐ True ☐ False
     False.
  3. The VC dimension of a parametric family of functions is always proportional to the number of parameters.
     ☐ True ☐ False
     False. e.g. a parameterized sine function
  4. There exists an infinite concept class $\mathcal{C}$ (i.e. $|\mathcal{C}| = \infty$) that is PAC learnable from itself.
     ☐ True ☐ False
     True. e.g. finite VC dimension classes

- Consider the learning problem introduced above. The notation $R(h) = \mathbb{E}_{x\sim\mathcal{D}} \mathbb{1}\left(c^*(x) \neq h(x)\right)$ denotes the population risk and $\hat{R}(h; S) = \frac{1}{|S|} \sum_{(x_i,y_i)\in S} \mathbb{1}\left(y_i \neq h(x_i)\right)$ the empirical risk, where $\mathbb{1}$ is the indicator function.

  Assume that $\mathcal{C} = \mathcal{H}$ and that the set $\mathcal{H}$ is finite. Write the optimization problem that corresponds to empirical risk minimization over $\mathcal{C}$. What is the minimum empirical risk that can be achieved for an arbitrary data set $S = \{(x_i, y_i)\}_{i=1}^n$?

  2 pts ☐

  .....

  .....

  .....

  .....

  .....

  Solution: $\min_{h\in\mathcal{C}} \hat{R}(h; S) = \min_{h\in\mathcal{C}} \frac{1}{|S|} \sum_{(x_i,y_i)\in S} \mathbb{1}\left(y_i \neq h(x_i)\right)$. Minimum empirical risk

is 0, achieved by $c^* \in \mathcal{C} = \mathcal{H}$.
Grading:

- $+1$ for the correct optimization problem for ERM.
- $+1$ for the correct mininum empirical risk.
- -0.5 for any redundant/incorrect statements, constraints, etc.

- For $\epsilon > 0$, let $\mathcal{C}_\epsilon = \{h \in \mathcal{C} : R(h) > \epsilon\}$ and assume that $\min_{h \in \mathcal{C}} \hat{R}(h; S) = 0$. Show that:

$$\mathbb{P}\left(\exists h \in \mathcal{C}_\epsilon \text{ s.t. } \hat{R}(h; S) = 0\right) \leq |\mathcal{C}_\epsilon|(1-\epsilon)^n$$

Hint: Use the union bound, i.e. for a set of events $\{E_1, ..., E_m\}$, $\mathbb{P}(E_1 \vee ... \vee E_m) \leq \sum_{i=1}^m \mathbb{P}(E_i)$.

3 pts

Solution: We know from the previous question the $r_{min} = 0$. $\mathbb{P}\left(\exists h \in \mathcal{C}_\epsilon \text{ s.t. } \hat{R}(h; S) = r_{min}\right) = \mathbb{P}\left(\bigvee_{h \in \mathcal{C}_\epsilon} \hat{R}(h; S) = 0\right) \leq \sum_{h \in \mathcal{C}_\epsilon} \mathbb{P}\left(\hat{R}(h; S) = 0\right) \leq |\mathcal{C}_\epsilon|(1-\epsilon)^n$.
Grading:

- $+1$ for transforming the given probability to the probability of the union.
- $+1$ for applying the union bound.
- $+1$ for bounding the probability of the union by $|\mathcal{C}_\epsilon|(1-\epsilon)^n$.
- -0.5 for any redundant/incorrect statements, constraints, etc.

- For $\epsilon > 0$ and $\delta > 0$, give a lower bound on n that is polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$ such that

$$\mathbb{P}\left(\exists h \in \mathcal{C} \text{ s.t. } \hat{R}(h; S) = 0 \text{ and } R(h) \leq \epsilon\right) \geq 1 - \delta. \tag{29}$$

Hint: You may also use the inequality $(1 - x)^z \leq e^{-xz}$ for small x.

3 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Solution: $\mathbb{P}\left(\exists h \in \mathcal{C}_\epsilon \text{ s.t. } \hat{R}(h; S) = r_{min}\right) \leq |\mathcal{C}_\epsilon|(1 - \epsilon)^n \leq |\mathcal{C}_\epsilon|e^{-\epsilon n}$. So $R(h) \leq \epsilon$ w.p. $1 - \delta \iff |\mathcal{C}_\epsilon|e^{-\epsilon n} \leq \delta$. Hence, we can write $n \geq \frac{1}{\epsilon}(\log |\mathcal{C}_\epsilon| + \log \frac{1}{\delta})$. This implies the polynomial bound: $n \geq \frac{1}{\epsilon}(\log |\mathcal{C}_\epsilon| + \frac{1}{\delta})$
Grading:

- $+1$ for bounding $|\mathcal{C}_\epsilon|(1 - \epsilon)^n$ by $|\mathcal{C}_\epsilon|e^{-\epsilon n}$.
- $+1$ for showing $R(h) \leq \epsilon$ w.p. $1 - \delta \iff |\mathcal{C}_\epsilon|e^{-\epsilon n} \leq \delta$.
- $+1$ for showing the lower bound on n.
- -0.5 for any redundant/incorrect statements, constraints, etc.
- $+3$ for the trivial solution, $i.e.$, the lower bound for n is any constant number. It is given that $\mathcal{C} = \mathcal{H}$, so the inequality (29) is always satisfied by setting $h = c^*$. NOTE: it needs to be a proof, so no points is given for answers that do not explain why ($e.g.$, just stating $n = 0$).

Question 11: Model selection (12 pts)

Each subquestion is independent of the others. You don't need to have solved the previous ones to solve the next one.

Prelude: The Kullback-Leibler divergence is a distortion measure between two distributions p and q over a finite sample space Θ defined as follows:

$$KL(p \parallel q) = \sum_{\theta \in \Theta} p(\theta) \log \left(\frac{p(\theta)}{q(\theta)} \right). \tag{30}$$

Demonstrate that

$$KL(p \parallel q) = -H[p] - \sum_{\theta \in \Theta} p(\theta) \log q(\theta), \tag{31}$$

where $H[p] = -\sum_{\theta \in \Theta} p(\theta) \log p(\theta)$ is the entropy of p:

3 pts ☐

Master solution:

$$KL(p \mid q) = \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right) = \sum_{x \in \mathcal{X}} p(x) \log p(x) + \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{1}{q(x)} \right)$$
$$= -(-\sum_{x \in \mathcal{X}} p(x) \log p(x)) - \sum_{x \in \mathcal{X}} p(x) \log q(x) = -H[p] - \sum_{x \in \mathcal{X}} p(x) \log q(x)$$

Note: Backward solution also accepted: going from equation 9 to 8.

Points distribution::

- Each step in master solution [1pt]
- Small error in placement of brackets in any of the equation [-0.5]

Setting: Let $\mathcal{X}$ be a space of possible datasets, Θ a **finite** space of hypotheses, and $R : \Theta \times \mathcal{X} \to \mathbb{R}$ a cost function. For $\theta \in \Theta$ and $X \in \mathcal{X}$, $R(\theta, X)$ measures how well the model θ fits the data X.

One of the main goals in machine learning is to compute from a given training set X' a *posterior distribution* $p(\cdot \mid X')$ over Θ. You may assume X' to be drawn from a distribution p^* over $\mathcal{X}$. For $\theta \in \Theta$, $p(\theta \mid X')$ measures our confidence that θ is the right hypothesis for X'.

In Lecture 2, we discussed two objectives that such a distribution should fulfil. In general, the two objectives cannot be fulfilled simultaneously, so a trade off is required.

Objective 1: *The posteriors should use the hypothesis class Θ uniformly when we average over all experiments.* We quantified this in the lecture by requiring $p(\cdot \mid X')$ to be as "close" as possible to the uniform distribution over Θ, in expectation with respect to X'. Formalize this objective, as done in the lecture, and then show that it is equivalent to requiring $p(\cdot \mid X')$, for $X' \in \mathcal{X}$, to maximize

$$\mathbb{E}_{X'}[H[p(\cdot \mid X')]], \tag{32}$$

where H is the entropy.

5

pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Master solution:
Formalise:
using minimisation of KL divergence between $p(\cdot \mid X')$ and uniform distribution over Θ

$$\arg\min_{\{p(\cdot|X')_{X' \in \mathcal{X}}\}} E_{X'}[KL(p(\cdot \mid X') \mid\mid Unif_\theta(\cdot))]$$

$$\arg\min_p E_{X'}(-H(p(\cdot \mid X')) - \sum_\theta p(\theta \mid X') \log \mid \theta \mid^{-1})$$

The right term is constant as $\log \mid \theta \mid^{-1}$is constant and $\sum_\theta p(\theta \mid X') = 1$

$$\arg\min_p -E_{X'}(H(p(\cdot \mid X'))$$

$$\arg\max_p E_{X'}(H(p(\cdot \mid X'))$$

Points distribution::

- Formalise by minimising KL divergence and writing KL divergence term correctly [2 pt]
- Applying formula for KL divergence for the given distributions [1.5 pt]
- Proof of constant term of right part [1 pt]
- maximizing to maximizing using minus term [0.5 pt]
- If maximizing instead of minimizing KL divergence mentioned in formalisation [-1.5 pt]

Objective 2: *The distribution must minimize the description length on test data.* We formalized this by requiring $p(\cdot \mid X)$ to minimize

$$\mathbb{E}_{X',X''}\left[\mathbb{E}_{\theta|X'}[-\log p(\theta \mid X'')]\right]. \tag{33}$$

Here, X'' is another dataset drawn from p^*. Show that we can bound this expression from below with

$$\mathbb{E}_{X',X''}\left[-\log \kappa(X', X'')\right], \tag{34}$$

where

$$\kappa(X', X'') = \sum_\theta p(\theta \mid X')p(\theta \mid X''). \tag{35}$$

2

pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Master solution:
Jensen's inequality: let f(x) be a convex function, then:

$$E_X f(X) \geq f[E_X]$$

Lower bound on the given objective (-log(X) is convex):

$$\mathbb{E}_{X',X''} \sum_{\theta} p(\theta \mid X')[-\log p(\theta \mid X'')] \geq \mathbb{E}_{X',X''}[-\log \sum_{\theta} p(\theta \mid X')p(\theta \mid X'')]$$

$$\mathbb{E}_{X',X''} \sum_{\theta} p(\theta \mid X')[-\log p(\theta \mid X'')] \geq \mathbb{E}_{X',X''}\left[-\log \kappa(X', X'')\right]$$

Points distribution::

- Mention of Jensen Shanon inequality or $-log$ convex [1pt]
- Expectation of $\mathbb{E}_{\theta|X'}$ to summation and using Jensen Shanon inequality (basically last two inequalities in master solution) [1pt]
- If inequality sign reversed [-1pt]

The new score: The lecture then demonstrated how to combine the two previous results into the following score:

$$\mathbb{E}_{X',X''}\left[\log\left(|\Theta|\,\kappa(X', X'')\right)\right] \tag{36}$$

Briefly explain what this score measures. Should it be maximized or minimized, why? You do not need to derive Formula 36 for this.

2 pts

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Master solution:
It quantifies how robust the algorithm is to fluctuations between two datasets X' and X'' drawn at random. Indeed, $\kappa(X', X'')$ is high only when $p(\cdot \mid X')$ and $p(\cdot \mid X'')$ look alike. Therefore, good learning algorithms shall maximize this score.
Keywords: measures generalization ability or robustness.
Note: referring to objective 1 and 2 for explaining what the score measures are also acceptable.

Points distribution::

- Maximize [1pt]
- Reason for maximization or what the score measures [1pt]