Advanced Machine Learning

Fall 2020

Prof. Joachim M. Buhmann

Final Exam

January 27, 2021

First and last name: _____

Student ID number: _____

Signature: _____

General Remarks

- Please check that you have all 30 pages of this exam.
- There are 100 points in total.
- The duration of the exam is 180 minutes.
- Do not spend too much time on a single question. You do not need 100 points to get 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. Later 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 receive 1 point per correct answer, -1 point per incorrect answer, and 0 points per unanswered question. If you would get less than 0 points in any block of true/false questions, you get 0 points instead.
- **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	Regression	12		
2	ML and MAP	7		
3	Bias and Variance	13		
4	Gaussian processes	13		
5	Linear discriminators	6		
6	Kernels	8		
7	Ranking SVM	7		
8	Ensembles	15		
9	ELBO	9		
10	PAC learning	10		
Total		100		

Question 1: Regression (12 pts)

Which statements are true and which are false?

5 pts

1) Ridge regression can be used to address the problem of multicollinearity.
   ☐ True ☐ False
   True

2) In ridge regression, the bias of the estimator increases with the regularization parameter.
   ☐ True ☐ False
   True

3) Compared with LASSO, ridge regression is more likely to end up with higher sparsity of coefficients.
   ☐ True ☐ False
   False

4) The cost function of LASSO is differentiable and convex.
   ☐ True ☐ False
   False

5) Ridge regression has a closed form solution for estimating the coefficients.
   ☐ True ☐ False
   True

Let $\{(x_1, y_1), ..., (x_n, y_n)\}$ be an i.i.d sample. Denote by $X \in \mathbb{R}^{n \times d}$ the matrix where the row i represents x_i, and by $Y \in \mathbb{R}^n$ the vector where entry i is y_i. Consider the loss function $L(x, y, \beta) = (y - x^\top \beta)^2$, where $x, \beta \in \mathbb{R}^d$ and $y \in \mathbb{R}$.

- Write down the expressions for the expected loss $\mathcal{R}(\beta)$ and the empirical loss $\hat{\mathcal{R}}(\beta)$.

  2 pts

  .....

  .....

  .....

  Expected loss: $\mathcal{R}(\beta) = \mathbb{E}_{x,y}[y - x^\top \beta]^2$.
  Empirical loss: $\hat{\mathcal{R}}(\beta) = \frac{1}{n}\sum_{i=1}^n (y_i - x_i^\top \beta)^2$ **OR** $\hat{\mathcal{R}}(\beta) = \frac{1}{n}(Y - X\beta)^\top (Y - X\beta)$
  Grading Scheme: Each Loss worths one point. Expected loss should clearly include either Expection mark $\mathbb{E}$ or integral mark $\int$. Empirical loss should be averaged loss of every single example from 1 to n. Without averaging ($\frac{1}{n}$) will cause -0.5. Typo will cause -0.5 point for each Loss. .

- Derive a closed-form expression for $\hat{\beta} = \arg\min_\beta \hat{\mathcal{R}}(\beta)$ in terms of X and Y.

  3 pts

  .....

  .....

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

$$\begin{aligned}\hat{\mathcal{R}}(\beta) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2 \\ &= \frac{1}{n}(Y - X\beta)^\top (Y - X\beta)\end{aligned} \tag{1}$$

To minimize $\hat{\mathcal{R}}(\beta)$ by β, let the gradient $\nabla_\beta \hat{\mathcal{R}}(\beta) = 0$.

$$\begin{aligned}\nabla_\beta \hat{\mathcal{R}}(\beta) &= -\frac{2}{n}X^\top (Y - X\beta) = 0 \\ \hat{\beta} &= (X^\top X)^{-1} X^\top Y\end{aligned} \tag{2}$$

Grading Scheme:
1 point: Correctly writing down the expression of $\hat{\mathcal{R}}(\beta)$ and letting the gradient equal to zero.
1 point: Correctly calculate the gradient $\nabla_\beta \hat{\mathcal{R}}(\beta) = 0$.
1 point: Correctly writing down the expression of $\hat{\beta}$ with X and Y.

- Assume now that the matrices X and Y are random. Furthermore, $Y = X\beta + \xi$, where $\xi \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}), \beta \in \mathbb{R}^d$. Calculate $\mathbb{E}_{\xi|X,Y}[\hat{\beta}]$, where $\hat{\beta} = \arg\min_\beta \hat{\mathcal{R}}(\beta)$. Show that $\hat{\beta}$ is an unbiased estimator of β.

2 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T Y \\ &= (X^T X)^{-1} X^T (X\beta + \xi) \\ &= \beta + (X^T X)^{-1} X^T \xi \end{aligned} \tag{3}$$

$$\begin{aligned} \mathbb{E}_{\xi|X,Y} &= \mathbb{E}[\beta + (X^T X)^{-1} X^T \xi] \\ &= \beta + (X^T X)^{-1} X^T \mathbb{E}[\xi] \end{aligned} \tag{4}$$

Because $\mathbb{E}[\xi] = 0$, we get $\mathbb{E}_{\xi|X,Y} = \beta$. Therefore, $\hat{\beta}$ is an unbiased estimator of β.

Grading Scheme:

1 point: Correctly writing down the expression of $\hat{\beta}$ and replacing Y with the given expression $Y = X\beta + \xi$.

1 point: Correctly writing down that the reasoning process of $\hat{\beta}$, which is the sum of β and $(X^T X)^{-1} X^T \mathbb{E}[\xi]$. Writing down $\mathbb{E}_{[}\hat{\beta}] = \beta$ because $\mathbb{E}[\xi] = 0$.

Question 2: ML and MAP Estimation (7 pts)

Let $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ denote an i.i.d. sample, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Assume a linear model $y_i = \beta^\top x_i + \epsilon$, where ϵ is drawn from $\mathcal{N}(0, \sigma^2)$. We want to estimate $\beta \in \mathbb{R}^d$.

- Write the log-likelihood $\log \mathcal{L}(\beta) = \log P(D|\beta)$, treating all terms that don't depend on β as constants.

4 pts

Noise ϵ being sampled from a Gaussian implies that the response y becomes a draw from $y \sim \mathcal{N}(\beta^\top x, \sigma^2)$. Thus the probability distribution of the response variable y is $P(y|x, \beta) \propto \mathcal{N}(\beta^\top x, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\beta^\top x)^2)}{2\sigma^2}\right]$. The log-likelihood is then given by

$$\begin{aligned}
\log P(y|x, \beta) &= \sum_{n=1}^{N} \log P(y_n|x_n, \beta) \\
&= \sum_{n=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_n - \beta^\top x)^2}{2\sigma^2}\right] \\
&= \sum_{n=1}^{N} \left\{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_n - \beta^\top x_n)^2}{2\sigma^2}\right\}.
\end{aligned}$$

(terms non dependent on β should be omitted from solution)

Grading Scheme:

2 points: $\log P(y, x|\beta) \propto \sum_{n=1}^{N} \log P(y_n|x_n, \beta)$

1 point: $\sum_{n=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(y_n-\beta^\top x)^2}{2\sigma^2}$

1 point: $C_1 \sum_{n=1}^{N} -(y_n - \beta^\top x_n)^2 + C_2$

- Assume a Laplacian prior with zero mean and scale parameter $\tau = \lambda^{-1}$ for each weight β_i for $i \leq d$. The *pdf* of such a Laplacian is:

$$f(\beta_i|0, \lambda^{-1}) = \frac{1}{2\lambda^{-1}} \exp(-\frac{|\beta_i|}{\lambda^{-1}}).$$

What type of regularization corresponds to performing MAP using this prior?

1 pts

☐ Elastic Net ☐ Lasso ☐ Ridge

Lasso

- In expectation, how does the MAP and the MLE compare with each other when the number of samples goes to infinity?

1 pts

. . . . . .

. . . . . .

For an infinite number of samples, the only contributing factor is the likelihood, so the MAP and MLE are equal

Grading Scheme:

1 point: $\beta_{MAP} \rightarrow \beta_{MLE}$

- What type of regularization corresponds to performing MAP using a Gaussian prior?

1 pts

☐ Elastic Net ☐ Lasso ☐ Ridge

Ridge

Question 3: Bias and Variance (13 pts)

1. List at least 2 disadvantages of the leave-one-out cross validation in comparison to the 10-fold cross validation?

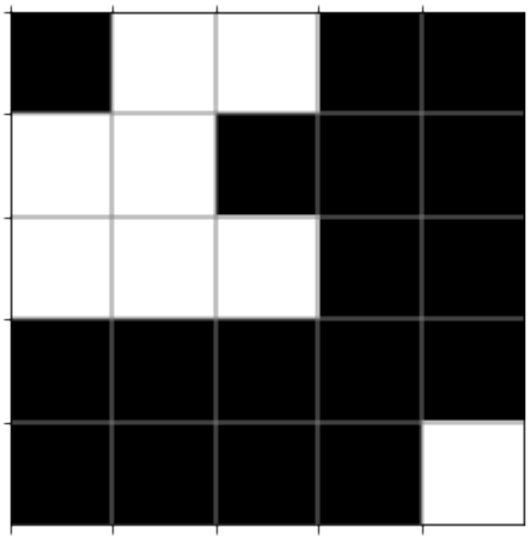
2 pts ☐

. . . . . .

. . . . . .

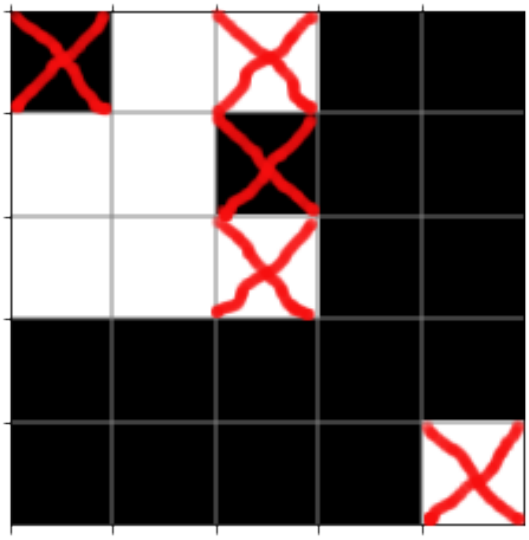
. . . . . .

2. In the figure below, each cell represents a point and its color (black or white) corresponds to the point's label. Neighbors are defined according to the Manhattan distance (the distance between two points measured along axes at right angles, e.g. in a plane with p_1 at (x_1, y_1) and p_2 at (x_2, y_2), it is $|x_1 - x_2| + |y_1 - y_2|$). What is the leave-one-out cross validation error of the 1NN classifier? If there is more than one nearest neighbor, use the majority voting. Mark the cells misclassified by the 1NN classifier.



2 pts ☐

1. Optimistic error estimation due to highly correlated training sets; computational cost.
Grading: 1pt each



2. $0.2 = \frac{5}{25}$, see Fig.3 Grading: 1pt for the correct picture, 1pt for the correct final answer

3. Which of the following approaches help to reduce the variance of the model being trained?

9 pts

1) Adding more features.
☐ True ☐ False
False

2) Increasing k when the model is trained with k-nearest neighbors.
☐ True ☐ False
True

3) L_2 regularization.
☐ True ☐ False
True

4) Putting a Laplace prior on the estimated parameters.
☐ True ☐ False
True

5) Using deeper decision trees.
☐ True ☐ False
False

6) Increasing the bias.
☐ True ☐ False
False

7) Dimensionality reduction.
☐ True ☐ False
True

8) Using more data.
☐ True ☐ False
True

9) Using wider neural networks.
☐ True ☐ False

False Graded as 1pt for any answer (and no answer) because of dis-ambiguity caused by double-decent effect.

Question 4: Gaussian Processes (13 pts)

Let $\{(x_1, y_1), \dots, (x_n, y_n)\} \subseteq \mathbb{R}^d \times \mathbb{R}$ be a dataset and let $X = (x_1, \dots, x_n)$ and $Y = (y_1, \dots, y_n)^\intercal$ be the feature matrix and the label vector respective. Let $k_\theta(x, x')$ be a parameterized Gaussian process kernel, where θ is a hyperparameter and denote by $K_\theta(X) = \{k_\theta(x_i, x_j)\}_{i,j \leq n}$ the covariance matrix induced by k_θ.

- Using the simple Gaussian prior with covariance $K_\theta(X)$: $f|X \sim \mathcal{N}(0, K_\theta(X))$, and observations $Y = f + \epsilon$ with Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$, derive the predictive distribution $p(Y|X)$.

3 pts ☐

Let us call f the noise-free observations. Then the prior reads $f|X \sim \mathcal{N}(0, K)$, and $Y = f + \epsilon$, then $Y|X$ is Gaussian with parameters $\mu = \mathbb{E}[f + \epsilon] = 0$ while $\mathbb{V}[f + \epsilon] =$

$\mathbb{E}[(f+\epsilon)^T(f+\epsilon)] = \mathbb{E}[ff^T + 2f\epsilon^t + \epsilon\epsilon^T] = K + \sigma^2 1$ for independence of ϵ and f.

- Compute the log likelihood $\mathcal{L}(\theta)$ as a function of the hyperparameters θ, dropping any irrelevant constant terms.

3 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

$\mathcal{L}(\theta) = \log p(Y|X;\theta) = \log \mathcal{N}(Y|0, K(\theta)+\sigma^2 1) = -1/2 y^T (K(\theta)+\sigma^2 1)^{-1} y - 1/2 \log |K(\theta)+\sigma^2 1|$

- Assume that $\theta \in \mathbb{R}^m$ and give an expression for computing $\frac{\partial \mathcal{L}}{\partial \theta_j}$, for $j \leq m$, in terms of $K_\theta(X)$, σ, $\frac{\partial K_\theta(X)}{\partial \theta_j}$. Use the formulas $\frac{\partial A^{-1}}{\partial \theta_j} = -A^{-1}\frac{\partial A}{\partial \theta_j}A^{-1}$ and $\frac{\partial \log|A|}{\partial \theta_j} = Tr(A^{-1}\frac{\partial A}{\partial \theta_j})$. Assuming that $(K_\theta(X))^{-1}$ and $\frac{\partial K_\theta(X)}{\partial \theta_j}$ have already been computed, calculate the time complexity of computing $\left(\frac{\partial \mathcal{L}}{\partial \theta_j}\right)_{j\leq m}$.

4 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

$\frac{\partial \mathcal{L}(\theta)}{\partial \theta_i} = -1/2y^T\frac{\partial (K(\theta)+\sigma^2 1)^{-1}}{\partial \theta_i}y - 1/2\frac{\partial \log|K(\theta)+\sigma^2 1|}{\partial \theta_i} = +1/2y^T K_\sigma^{-1}\frac{\partial K(\theta)}{\partial \theta_i}K_\sigma^{-1}y - 1/2Tr(K_\sigma^{-1}\frac{\partial K(\theta)}{\partial \theta_i})$. The cost of the gradient is $O(mn^2)$.

- Write a gradient descent algorithm to maximize the log-likelihood.

3 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

initialize θ_0. define a learning rate η. for any $k > 0$, $\theta_k = \theta_{k-1} - \eta\nabla\mathcal{L}(\theta_{k-1})$, until convergence

Question 5: Linear discriminators (6 pts)

Which statements are true?

6 pts

1) The perceptron algorithm converges to the same solution for any initialization.
   ☐ True ☐ False
   False.

2) The classifier in the perceptron algorithm is linear.
   ☐ True ☐ False
   True.

3) The standard perceptron algorithm always converges for any training data set.
   ☐ True ☐ False
   False.

4) A linear model can be trained only with a symmetric loss function: $L(\mathbf{w}^\top x - y) = L(-(\mathbf{w}^\top x - y))$.
   ☐ True ☐ False
   False.

5) The logistic loss $\log(1 + e^{-t})$ and the hinge loss functions $\max\{0, 1 - t\}$ have the same asymptotics for large deviations from zero ($t \to \infty$ or $t \to -\infty$).
   ☐ True ☐ False
   True.

6) The exponential loss function e^t is more robust to outliers than the hinge loss.
   ☐ True ☐ False
   False.

Question 6: Kernels (8 pts)

Consider a SVM trained on $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ with a kernel $K(\cdot, \cdot)$, where only m of them ($m < n$) are support vectors.

What is the time complexity of making a prediction for a new point x, as a function of n and m? (Assume $K(\cdot, \cdot)$ can be computed in constant time for any point.)

3 pts

$O(m)$ or $\Theta(m)$.

Formally prove that the Gram matrix XX^T is a kernel matrix, that is, symmetric and positive semi-definite, for any $X \in \mathbb{R}^{n \times d}$.

3 pts

. . . . . .

. . . . . .

. . . . . .

$\sum_i \sum_j < x_i, x_j > a_i a_j = \sum_i \sum_j < a_i x_i, a_j x_j >=< \sum_i a_i x_i, \sum_j a_j x_j >= \| \sum_i a_i x_i \|^2 >= 0$ for any $a_1, \ldots, a_n$, $x_1, \ldots, x_n$.

Are the following statements true or false?

2 pts ☐

1) If K is an RBF kernel, $K(x', x'') = \exp(-\gamma \|x' - x''\|^2)$, one can find a parameter $\gamma > 0$ such that the SVM classifier with this kernel is equivalent to the nearest neighbor classifier.

☐ True ☐ False

False.

2) Matrix $\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$ is a kernel matrix.

☐ True ☐ False

True.

Question 7: Ranking SVM (7 pts)

Let x_i^j be a set of training examples where $1 \leq j \leq k$ denotes the class number and $1 \leq i \leq n_j$ the index within each class. There are n_j samples in class j. We assume that the classes can be ordered in a similar fashion to Figure 1. Consider the problem of fitting a set of $k-1$ parallel hyperplanes $((\mathbf{w}, b_1), ..., (\mathbf{w}, b_{k-1}))$. The hyperplane $(\mathbf{w}, b_j)$ should have maximum margin and separate classes j and $j+1$. ϵ_i^j and γ_i^{j+1} are the slack variables for the two classes j and $j+1$.

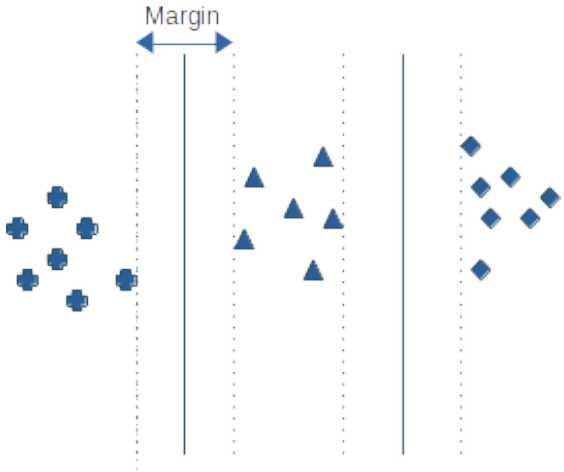


Figure 1

$$\min_{\mathbf{w}, b_j, \epsilon_i^j, \gamma_i^{j+1}} \quad \frac{1}{2}||\mathbf{w}||^2 + C \sum_i \sum_j \left(\epsilon_i^j + \gamma_i^{j+1}\right) \tag{5}$$

$$\text{s.t. } \mathbf{w}^\top x_i^j + b_j \leq -1 + \epsilon_i^j$$
$$\mathbf{w}^\top x_i^{j+1} + b_j \geq 1 - \gamma_i^{j+1}$$
$$\epsilon_i^j \geq 0, \quad \gamma_i^{j+1} \geq 0$$

1. Formulate the Lagrangian for the constrained optimization problem above

5 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

2. State the conditions that the Lagrangian must satisfy at the optimal solution. You do not need to explicitly compute any derivatives.

2 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

1. The Lagrangian of the constrained optimization problem is given as follows:

$$L(\mathbf{w}, \epsilon, \mathbf{b}, \alpha, \beta, \xi, \lambda) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i,j}(\epsilon_i^j + \gamma_i^{j+1}) + \sum_{i,j}\alpha_i^j(\mathbf{w}x_i^j + b_j + 1 - \epsilon_i^j) + \sum_{i,j}\beta_i^j(1 - \gamma_i^{j+1} - b_j - \mathbf{w}x_i^{j+1}) - \sum_{i,j}\xi_i^j\epsilon_i^j - \sum_{i,j}\lambda_i^{j+1}\gamma_i^{j+1}$$

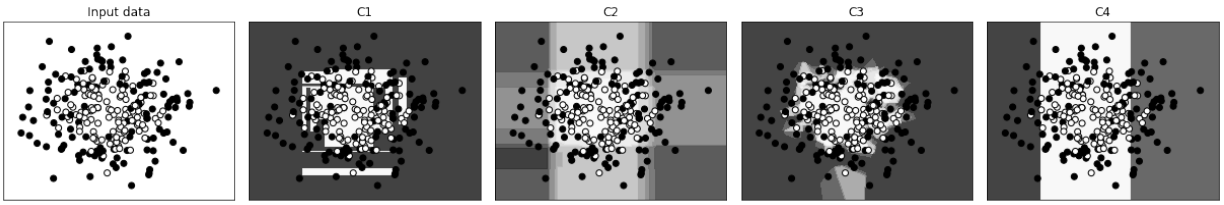
2. The stationarity conditions are given as follows:

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \rightarrow ...$$
$$\frac{\partial L}{\partial \mathbf{b}} = 0 \rightarrow ...$$
$$\frac{\partial L}{\partial \epsilon_i^j} = 0 \rightarrow ...$$
$$\frac{\partial L}{\partial \gamma_i^{j+1}} = 0 \rightarrow ...$$

Question 8: Ensembles (15 pts)

1) The following figure depicts a dataset for binary classification and four ensembles trained on a same training set. Match the classifier name to its corresponding plot ($C1, C2, C3, C4$).

4 pts



- AdaBoost with deep decision trees ...
- Bagging shallow decision trees ...
- Shallow decision tree ...
- Bagging 1NN classifiers ...

- AdaBoost with deep decision trees C1
- Bagging shallow decision trees C2
- Shallow decision tree C4
- Bagging 1NN classifiers C3

2) Let X and Y be two random variables taking values in $\mathbb{R}^D$ and $\{-1, 1\}$, respectively. Consider the binomial deviance loss function, defined as

$$\mathcal{L}(y, f(x)) = \log\left(1 + e^{-2yf(x)}\right), \tag{6}$$

where $f : \mathbb{R}^D \to \mathbb{R}$ is a regression function.

a) Derive the formula for the expected value $\mathbb{E}_{Y|x}\left[L\left(Y, f(x)\right)\right]$ in terms of $P(Y = 1 \mid X = x)$ and $P(Y = -1 \mid X = x)$.

3 pts

.....

.....

.....

.....

b) Find the derivative of $\mathbb{E}_{Y|x}\left[L\left(Y, f(x)\right)\right]$ with respect to $f(x)$.

3 pts

c) Derive the population minimizer $f^*(x) = \arg\min_{f(x)} \mathbb{E}_{Y|x}\left[L\left(Y, f(x)\right)\right]$.

5 pts

a) Derive the formula for the expected value (3pt)

$$
\begin{aligned}
E[f(x)] &:= \mathbb{E}_{Y|x}\left[L\left(Y, f(x)\right)\right] \\
&= \mathbb{E}_{Y|x}\left[\log\left(1 + e^{-2Y f(x)}\right)\right] \\
&= P\left(Y = 1 \mid X = x\right)\left(\log\left(1 + e^{-2f(x)}\right)\right) + P\left(Y = -1 \mid X = x\right)\left(\log\left(1 + e^{2f(x)}\right)\right).
\end{aligned}
$$

Points distribution:

- 1pt for each part (Y=1, Y=-1) correct

– 1pt for correct summation

b) Find the derivative w.r.t. $f(x)$ (3pt)

$$\begin{aligned}\frac{\partial E[f(x)]}{\partial f(x)} =& \mathbf{P}\left(Y=1 \mid X=x\right)\left(\frac{1}{1+e^{-2f(x)}}e^{-2f(x)}\cdot(-2)\right)\\ &+\mathbf{P}\left(Y=-1 \mid X=x\right)\left(\frac{1}{1+e^{-2f(x)}}e^{2f(x)}\cdot 2\right).\end{aligned}$$

Points distribution:

– 1pt for a correct derivative of log(...)

– 1pt for correct treatment of the two parts, i.e. sign between the two parts $(+)$, not dropping $\mathbf{P}\left(Y=1 \mid X=x\right), \mathbf{P}\left(Y=-1 \mid X=x\right)$

– 1pt for correct computation (doesn't need to be simplified)

Comments:

– -1pt for including $\frac{\partial f(x)}{\partial x}$

– no subtraction for follow-up error from part a, if the formula for the expected value somewhat valid, otherwise -1pt or 0pt total

c) Setting $\frac{\partial E[f(x)]}{\partial f(x)} = 0$ (to find argmin) gives

$$\begin{aligned}0 &= 2\left(\mathbf{P}\left(Y=-1 \mid X=x\right)\frac{e^{2f(x)}}{1+e^{2f(x)}} - \mathbf{P}\left(Y=1 \mid X=x\right)\frac{e^{-2f(x)}}{1+e^{-2f(x)}}\right)\\ 0 &= 2\left(\mathbf{P}\left(Y=-1 \mid X=x\right)\frac{e^{2f(x)}}{1+e^{2f(x)}} - \mathbf{P}\left(Y=1 \mid X=x\right)\frac{1}{1+e^{2f(x)}}\right)\\ 0 &= \mathbf{P}\left(Y=-1 \mid X=x\right)e^{2f(x)} - \mathbf{P}\left(Y=1 \mid X=x\right)\\ e^{2f(x)} &= \frac{\mathbf{P}\left(Y=1 \mid X=x\right)}{\mathbf{P}\left(Y=-1 \mid X=x\right)}\\ f(x) &= \frac{1}{2}\log\left(\frac{\mathbf{P}\left(Y=1 \mid X=x\right)}{\mathbf{P}\left(Y=-1 \mid X=x\right)}\right).\end{aligned}$$

Points distribution:

– 1pt for "setting derivative to 0"

– 1pt for correct initial computations

– 1pt for arriving to the following point (or equivalent):

$$\frac{e^{2f(x)}+1}{e^{-2f(x)}+1} = \frac{\mathbf{P}\left(Y=1 \mid X=x\right)}{\mathbf{P}\left(Y=-1 \mid X=x\right)}$$

– 1pt for arriving to the following point (or equivalent):

$$e^{2f(x)} = \frac{\mathbf{P}\left(Y=1 \mid X=x\right)}{\mathbf{P}\left(Y=-1 \mid X=x\right)}$$

– 1pt for correct final answer

Comments:

- no points are subtracted if there is no mention about the denominator$\neq 0$, since it was also not explicitly written in the master solution of the corresponding exercise in the exercise sheet (tutorial)
- lack of derivations but correct answer gives 3pt (if mentioned about setting the derivative to 0), since it was explicitly written "derive" in the task description
- for full derivation but with some major mistake (such as incorrect use of log), influencing the difficulty of computations, leading to wrong result: 2pt
- for full derivation but with some minor mistake (such as wrong sign, simplifications), influencing the difficulty of computations, leading to wrong result: 3pt
- for full derivation but with some minor mistake not influencing the difficulty of computations: 4/4.5pt
- no subtraction for follow-up errors, unless leading to substantial simplification of computation (then -1pt) or showing misunderstanding of the concepts (then only 0/1pt awarded)
- -0.5pt for mistakes that are clearly typos

Question 9: Evidence lower bound (9 pts)

Recall the notation from the lecture:

- $p_{\theta'}(\cdot)$ is a distribution over a representation space $\mathcal{Z}$, parametrized by $\theta' \in \Theta'$.
- For $z \in \mathcal{Z}$, $p_\theta(\cdot|z)$ is a conditional distribution over a measurement space $\mathcal{X}$, parametrized by $\theta \in \Theta$.
- For $x \in \mathcal{X}$, $q_\phi(\cdot|x)$ is a tractable distribution over $\mathcal{Z}$, parametrized by $\phi \in \Phi$. It is intended to approximate $p_{\theta,\theta'}(\cdot|x) \propto p_{\theta'}(\cdot)p_\theta(x|\cdot)$.

Let $x_1, \ldots, x_m \subseteq \mathcal{X}$. In the lecture we demonstrated that:

$$\sum_{i \leq n} \log(p_{\theta,\theta'}(x_i)) \geq \sum_{i \leq n} \mathsf{elbo}_{\theta',\theta,\phi}(x_i), \tag{7}$$

where

$$\mathsf{elbo}_{\theta',\theta,\phi}(x_i) := \mathbb{E}_{z \sim q_\phi(\cdot|x_i)}[\log(p_\theta(x_i|z))] + \mathbb{E}_{z \sim q_\phi(\cdot|x_i)}\left[\log\left(\frac{p_{\theta'}(z)}{q_\phi(z|x_i)}\right)\right] \tag{8}$$

Demonstrate that, for fixed $\theta' \in \Theta', \theta \in \Theta$:

$$\arg\min_{\phi \in \Phi}\left\{\mathbb{E}_{z \sim q_\phi(\cdot|x_i)}\left[\log\left(\frac{q_\phi(z|x_i)}{p_{\theta,\theta'}\left(z \mid x_i\right)}\right)\right]\right\} = \arg\max_{\phi \in \Phi}\{\mathsf{elbo}_{\theta',\theta,\phi}(x_i)\} \tag{9}$$

Solution: transoform expression for elbo:

$$\text{elbo}_{\theta',\theta,\phi}(x_i) := \mathbb{E}_{z\sim q_\phi(\cdot|x_i)}[\log(p_\theta(x_i|z))] + \mathbb{E}_{z\sim q_\phi(\cdot|x_i)}\left[\log\left(\frac{p_{\theta'}(z)}{q_\phi(z|x_i)}\right)\right] = \quad (10)$$

$$= \mathbb{E}_{z\sim q_\phi(\cdot|x_i)}[\log(p_\theta(x_i|z))] + \mathbb{E}_{z\sim q_\phi(\cdot|x_i)}[\log(p_{\theta'}(z))] - \mathbb{E}_{z\sim q_\phi(\cdot|x_i)}[\log(q_\phi(z|x_i))] = \quad (11)$$

$$= \mathbb{E}_{z\sim q_\phi(\cdot|x_i)}[\log(p_\theta(x_i|z)p_{\theta'}(z))] - \mathbb{E}_{z\sim q_\phi(\cdot|x_i)}[\log(q_\phi(z|x_i))] = \quad (12)$$

$$= \mathbb{E}_{z\sim q_\phi(\cdot|x_i)}[\log(p_{\theta,\theta'}(x_i,z)] - \mathbb{E}_{z\sim q_\phi(\cdot|x_i)}[\log(q_\phi(z|x_i))] = \quad (13)$$

$$= \mathbb{E}_{z\sim q_\phi(\cdot|x_i)}[\log(p_{\theta,\theta'}(z|x_i)] + \mathbb{E}_{z\sim q_\phi(\cdot|x_i)}[\log(p_{\theta,\theta'}(x_i)] - \mathbb{E}_{z\sim q_\phi(\cdot|x_i)}[\log(q_\phi(z|x_i))] = \quad (14)$$

$$= -\mathbb{E}_{z\sim q_\phi(\cdot|x_i)}\left[\log\left(\frac{q_\phi(z|x_i)}{p_{\theta,\theta'}(z|x_i)}\right)\right] + \mathbb{E}_{z\sim q_\phi(\cdot|x_i)}[\log(p_{\theta,\theta'}(x_i)] = \quad (15)$$

$$(16)$$

Because the term $\mathbb{E}_{z\sim q_\phi(\cdot|x_i)}[\log(p_{\theta,\theta'}(x_i)]$ doesn't depend on ϕ we have:

$$\arg\min_{\phi\in\Phi}\left\{\mathbb{E}_{z\sim q_\phi(\cdot|x_i)}\left[\log\left(\frac{p_{\theta'}(z)}{q_\phi(z|x_i)}\right)\right]\right\} = \arg\min_{\phi\in\Phi}\{-\text{elbo}_{\theta',\theta,\phi}(x_i)\} = \quad (17)$$

$$= \arg\max_{\phi\in\Phi}\{\text{elbo}_{\theta',\theta,\phi}(x_i)\} \quad (18)$$

Q.E.D
Points distribution:

- Clear solution:
  9 pts – everything is correct or a few small typos;
  8 pts – everything correct but minor explanation are unclear or a lot of typos;
- Unclear solution:
  (Comment: there is some logic, but the student did mistakes by omitting terms in the non-obvious way or explanation statements are quite fuzzy etc.):
  2 pts – if students made some mistakes at the beginning but the further logic exists;
  2 pts - if there is some correct conceptual description but no math involved (normally, students write about elbo, infomax, KL);
  If wrong/incorrect statements start later in the solution then I added points to basic 0 pts:
  +2 pts if probability product rule is applied correctly;
  +3 pts if logarithms are unrolled correctly;
  So, it is possible to get 0,2,3 or 5 pts here.
- Absent solution:
  0 pts – no solution or only 1-2 computational transformation without further solution/logic/explanation or completely incorrect math.
- Attention!
  I didn't penalize points for "-" before "argmax", because this mistake was done by me in the condition of the question, and despite the announcement at the exam it could be very confusing.

Question 10: PAC Learning: Subset Learning (10 pts)

Let $D = \{1, ..., d\}$ be a set with fixed $d \in \mathbb{N}$. Let the instance space $\mathcal{X}$ consist of all subsets of D, i.e. $\mathcal{X} = \{A \mid A \subseteq D\}$. For a subset A of D, let $c_A : \mathcal{X} \to \{0, 1\}$ be the function such that, for $X \in \mathcal{X}$,

$$c_A(X) = \begin{cases} 1, & \text{if } A \subseteq X, \\ 0, & \text{otherwise.} \end{cases}$$

Let $\mathcal{C} = \{c_A \mid A \subseteq D\}$.

- Consider $d = 10$, $A = \{4, 6, 7\}$. Compute the following:

1 pts

$c_A(\{2, 4, 5, 6, 7\}) =$
$c_A(\{1, 6, 7, 9\}) =$
$c_A(\{3, 4, 5, 6\}) =$
$c_A(\{2, 4, 5, 6, 7\}) = 1, c_A(\{1, 6, 7, 9\}) = 0, c_A(\{3, 4, 5, 6\}) = 0.$

Schema:

- 1 point if correct for each element of Q.

- Let $\hat{c}_n^*$ be an empirical risk minimizer. Recall the VC inequality:

$$\mathbf{P}\left(\mathcal{R}(\hat{c}_n^*) - \inf_{c \in \mathcal{C}} \mathcal{R}(c) > \epsilon\right) \leq 2|\mathcal{C}| \exp\left(-\frac{n\epsilon^2}{2}\right),$$

for all $\epsilon > 0$, n, and for $|\mathcal{C}| \leq N$. Let $\delta > 0$. Give a lower bound for n such that

$$\mathbf{P}\left(\mathcal{R}(\hat{c}_n^*) - \inf_{c \in \mathcal{C}} \mathcal{R}(c) > \epsilon\right) \leq \delta$$

holds and that is polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$.

3 pts

Choose n s.t. $2\left|\mathcal{C}\right|\exp\left(-\frac{n\epsilon^2}{2}\right) \leq \delta$. $n \geq \frac{2}{\epsilon^2}\log\left(\frac{2|\mathcal{C}|}{\delta}\right)$. If $n \geq \frac{2}{\epsilon^2}\frac{2|\mathcal{C}|}{\delta}$, then $n \geq \frac{2}{\epsilon^2}\log\left(\frac{2|\mathcal{C}|}{\delta}\right)$. Note, that $\frac{2}{\epsilon^2}\frac{2|\mathcal{C}|}{\delta}$ is polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$.

Schema:

- 2 points for correctly deriving $n \geq \frac{2}{\epsilon^2}\log\left(\frac{2|\mathcal{C}|}{\delta}\right)$.
- 1 point for deriving a sample size polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$.

- You may assume that there is no noise in observed labels, i.e. you may assume that $\inf_{c \in \mathcal{C}} \mathcal{R}(c) = 0$. Recall the definition of PAC learnability: $\mathcal{C}$ is PAC learnable from $\mathcal{H}$ if there is an algorithm $\mathcal{A}$ s.t. for all $0 < \epsilon, \delta < \frac{1}{2}$, for any distribution on $\mathcal{X}$, for any $c \in \mathcal{C}$, there is a polynomial function $poly(\cdot, \cdot, \cdot)$ s.t. if $\mathcal{A}$ receives a sample $\mathcal{Z}$ of size $n \geq poly\left(\frac{1}{\epsilon}, \frac{1}{\delta}, size(c)\right)$, then $\mathcal{A}$ outputs $\hat{c}$ such that $\mathbf{P}\left(\mathcal{R}\left(\hat{c}\right) \leq \epsilon\right) \geq 1 - \delta$. You may assume that $size(c) = 1$ for all $c \in \mathcal{C}$.

  Assume that there is an algorithm $\mathcal{A}$ that produces an ERM $\hat{c}_n^*$ given a sample $\mathcal{Z} = \{(X_i, y_i) : i \leq n\} \subseteq \mathcal{X} \times \{0, 1\}$ with $y_i = c(X_i)$ for some $c \in \mathcal{C}$. Show that $\mathcal{C}$ is PAC learnable from itself.

  4 pts ☐

  .....

  .....

  .....

  .....

  .....

  We have shown above that for $n \geq \frac{2}{\epsilon^2}\frac{2|\mathcal{C}|}{\delta}$, $\mathbf{P}\left(\mathcal{R}\left(\hat{c}_n^*\right) > \epsilon\right) \leq \delta$, i.e. $\mathbf{P}\left(\mathcal{R}\left(\hat{c}_n^*\right) \leq \epsilon\right) \geq 1 - \delta$. Given that $\frac{2}{\epsilon^2}\frac{2|\mathcal{C}|}{\delta}$ is polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$ and that there exists an algorithm which returns an ERM, $\mathcal{C}$ is PAC learnable from itself.

  Schema:

  - 1 points for deriving that for $n \geq \frac{2}{\epsilon^2}\frac{2|\mathcal{C}|}{\delta}$, $\mathbf{P}\left(\mathcal{R}\left(\hat{c}_n^*\right) \leq \epsilon\right) \geq 1 - \delta$.
  - 2 points for noting that there exists an algorithm for ERM and that n is polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$.
  - 1 point for concluding that $\mathcal{C}$ is PAC learnable, given everything above.

- Consider the following algorithm $\mathcal{A}$ for learning a hypothesis $\hat{c}_n \in \mathcal{C}$ from the data $\mathcal{Z} = \{(X_i, y_i) : 1 \leq i \leq n\} \subset \mathcal{X} \times \{0, 1\}$:

  1. $\hat{A} \leftarrow \{1, 2, ..., d\}$.
  2. **For** $i = 1, ..., n$ **do**:

     (a) **If** ..... **then**

     $\hat{A} \leftarrow$ .....

  3. $\hat{c}_n \leftarrow c_{\hat{A}}$.
  4. **Return** $\hat{c}_n$.

  Complete the missing lines so that $\mathcal{A}$ minimizes the empirical risk.

2 pts

If $y_i = 1$ **then** $\hat{A} \leftarrow \hat{A} \cap X_i$. In this way, if a feature is not in at least one positive instance, it will not be included in $\hat{A}$. Thus, $\hat{c}_n$ is consistent with the training data, i.e. $\hat{\mathcal{R}}_n(\hat{c}_n) = 0$; and hence, $\hat{c}_n$ is an ERM.

Schema:

- 2 points for filling in the missing lines of the pseudocode.