Prof. Joachim M. Buhmann

Final Exam

January 22, 2019

First and Last name: ______________________

Student ID (Legi) Nr: ______________________

Signature: ______________________

General Remarks

- Please check that you have all 20 pages of this exam.
- There are 100 points in total, and the exam duration is 180 minutes.
- Don't spend too much time on a single question. The maximum number of points is not required for the best grade.
- Remove all material from your desk which is not permitted by the examination regulations.
- 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 of the rector's office with a possible exclusion from the examination and it can have judicial consequences.
- Use a **black** or a **blue** pen to answer the questions. Don't use pencils.
- Provide only one solution to each exercise. Cancel invalid solutions clearly.
- **Grading of multiple choice questions**: Unless stated otherwise, you get 1 point per correct answer, -1 point per incorrect answer, but you cannot get less than zero points in any block of multiple choice questions. Multiple correct answers are possible.

Grade: ....

	Topic	Max. Points	Points Achieved	Checked
1	Regression	6		
2	Maximum Likelihood Estimation	12		
3	Cross-Validation	5		
4	Lasso Regression / Cross-Validation	4		
5	Generalization	6		
6	Feature Transformations	8		
7	Properties of Perceptrons	5		
8	Information Criteria	4		
9	Numerical Estimation	5		
10	Bagging	4		
11	AdaBoost	6		
12	SVMs and Kernels	13		
13	Structural SVMs	11		
14	PAC-Learning	11		
Total		100		

Question 1: Regression (6 pts)

Which of the following claims are true/false?

6 pts ☐

1) In Lasso regression, the regularizer can increase the sparsity of the resulting solutions.
☐ True ☐ False

2) The objectives of Ridge and Lasso regression are both convex and have closed-form solutions.
☐ True ☐ False

3) For regression problems, the least squares estimate of the model coefficients β has the smallest variance among all linear estimates.
☐ True ☐ False

4) Linear regression models can capture non-linear relationships between the input $\mathbf{x}$ and the response y when combined with suitable feature transformations.
☐ True ☐ False

5) From a Bayesian point of view, the regularizer in Lasso regression corresponds to a Laplace prior on the coefficients β.
☐ True ☐ False

6) Using the regularizer $\lambda\beta^\top\beta$ in Ridge regression amounts to increasing all eigenvalues of $\mathbf{X}^\top\mathbf{X}$ by $\sqrt{\lambda}$, and thus increases the stability of the optimization problem.
☐ True ☐ False

Question 2: Maximum Likelihood Estimation (12 pts)

1) Which of the following claims are true/false?

4 pts ☐

- Maximum likelihood estimation cannot be used for discrete random variables.
☐ True ☐ False

- Maximum likelihood estimators are always unbiased.
☐ True ☐ False

- Maximum likelihood estimators are consistent under suitable regularity assumptions.
☐ True ☐ False

- If $\hat{\theta}_{MLE}$ is the maximum likelihood estimate of θ, then $g(\hat{\theta}_{MLE})$ is the maximum likelihood estimate of $g(\theta)$ for any invertible function g.
☐ True ☐ False

2) Assume that $X_1, \dots, X_n$ are i.i.d. random variables distributed according to a normal distribution $\mathcal{N}(\mu, \sigma^2)$. Consider the following estimator $\hat{\mu_1}$ of the mean μ:

$$\hat{\mu_1} = \frac{1}{2}(X_1 + X_2)$$

Which of the following statements are true/false? **2 pts** ☐

- $\hat{\mu_1}$ is unbiased.
  ☐ True ☐ False

- $\hat{\mu_1}$ is consistent.
  ☐ True ☐ False

3) Assume again that $X_1, \dots, X_n$ are i.i.d. random variables distributed according to a normal distribution $\mathcal{N}(\mu, \sigma^2)$. Consider the following estimator $\hat{\mu_2}$ of the mean μ:

$$\hat{\mu_2} = \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n^2}$$

Which of the following statements are true/false? **2 pts** ☐

- $\hat{\mu_2}$ is unbiased.
  ☐ True ☐ False

- $\hat{\mu_2}$ is consistent.
  ☐ True ☐ False

4) Assume that $X_1, \dots, X_n$ are i.i.d. random variables whose probability density is given by

$$p(x|\theta) = \theta x^{\theta-1} \quad \text{for} \quad 0 < x < 1$$

for some $\theta > 0$. Write down the corresponding log-likelihood function $\mathcal{L}(\theta)$ for given data $\mathbf{x} = (x_1, \dots, x_n)$.

2 pts ☐

....

....

5) In the situation of the previous question, compute the maximum likelihood estimate $\hat{\theta}_{MLE}$ of θ for given data $\mathbf{x} = (x_1, \dots, x_n)$.

2 pts ☐

....

....

Question 3: Cross-Validation (5 pts)

The plot in Figure 1 shows 7 data points $x_i \in \mathbb{R}^2$ with labels $y_i \in \{0, 1\}$ (the points with $y_i = 0$ are represented by o, those with $y_i = 1$ by x):

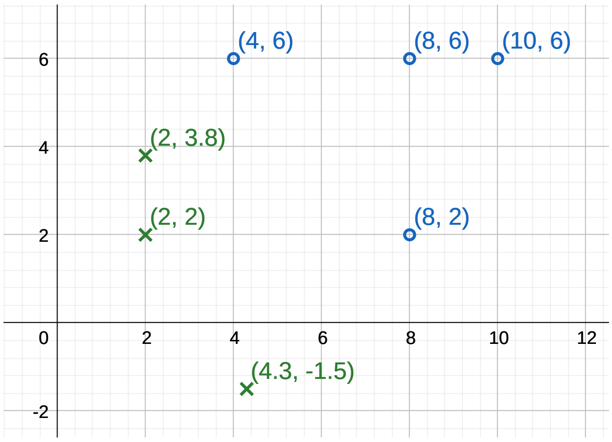


Figure 1: Labeled points from two classes.

Assume that we train a 1-nearest neighbor classifier (1-NN) c_ρ with respect to a distance function ρ on $\mathbb{R}^2$ on this dataset (recall that c_ρ assigns to a point $x \in \mathbb{R}^2$ the label of the training point which is closest with respect to ρ). We now want use leave-one-out cross-validation (LOOCV) to estimate the prediction error of c_ρ with respect to the loss function $Q(y, \hat{y}) = (y - \hat{y})^2$.

1. Compute the LOOCV estimate of the prediction error of c_ρ assuming that ρ is the Euclidean distance.

   2 pts

   ....

   ....

2. Compute the LOOCV functional of the prediction error of c_ρ assuming that ρ is the distance function given by $\rho((x_1, x_2), (x_1', x_2')) = \max\{|x_1 - x_1'|, |x_2 - x_2'|\}$.

   2 pts

   ....

   ....

3. In general, how many times do you need to fit a classifier to perform k-fold cross-validation on a dataset with n data points (where $n \geq k > 0$)?

   1 pts

   ....

Question 4: Lasso Regression / Cross-Validation **(4 pts)**

Assume that we want to train Lasso regression with gradient descent. Recall that the Lasso objective is

$$\min_{\beta} \frac{1}{N}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1,$$

where $X \in \mathbb{R}^{N\times p}$ and $y \in \mathbb{R}^N$. Assume that we want to tune λ with cross validation. Let $X^{(1)}_{train}$, $X^{(1)}_{test}$ be the training and test data for the split we are considering, and let $y^{(1)}_{train}$, $y^{(1)}_{train}$ be the corresponding responses. $X^{(1)}_{train}$ contains N_{train} points, while $X^{(1)}_{test}$ contains N_{test} points.

1. Compute the update β_{t+1} of the current solution β_t obtained by taking one step of gradient descent with stepsize γ_t, assuming that all components of $\beta_t \in \mathbb{R}^p$ are non-zero.

   2 pts ☐

   ....

   ....

2. After n steps of gradient descent we obtain a solution $\hat{\beta}_n \in \mathbb{R}^p$. Write down the expression for the estimate of the corresponding prediction error.

   2 pts ☐

   ....

   ....

Question 5: Generalization **(6 pts)**

Consider the two datasets X_{train} and X_{test} shown in Figure 2. Both contain 40 samples $(x, y) \in \mathbb{R}^2 \times \{-1, +1\}$. Assume that we have trained a classifier $c_{\hat{\theta}}$ on the training set $X_{train} = \{(x_1, y_1), \ldots, (x_{40}, y_{40})\}$ by solving

$$\hat{\theta} \in \arg\min_{\theta} \frac{1}{40}\sum_{i=1}^{40} Q(y_i, c_\theta(x_i)) + \lambda\|\theta\|_2^2.$$

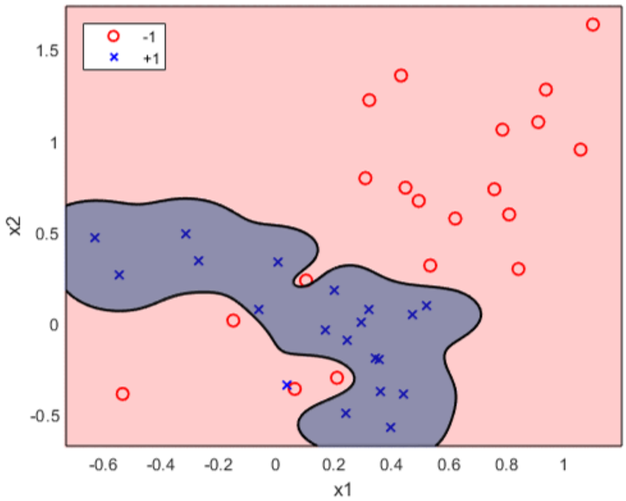
Here Q is a loss function, $\|\theta\|_2^2$ is a regularization term and λ is a hyperparameter that determines the strength of the regularization.

1. Compute the training error $\hat{R}(c_{\hat{\theta}}, X_{train})$ and the test error $\hat{R}(c_{\hat{\theta}}, X_{test})$, assuming that the loss function Q is the 0-1-loss, i.e., $Q(y_i, c(x_i)) = 0$ if $y_i = c(x_i)$, and $Q(y_i, c(x_i)) = 1$ if $y_i \neq c(x_i)$.

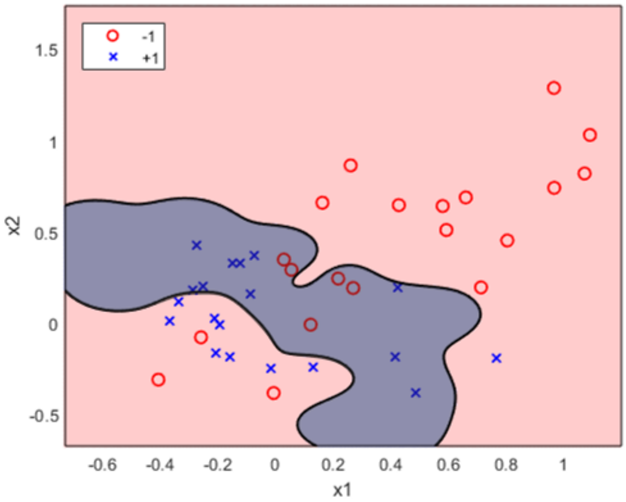
   2 pts ☐

   ....

   ....



(a) X_{train}



(b) X_{test}

Figure 2: Datasets X_{train} and X_{test}. The classifier $c_{\hat{\theta}}$ predicts $+1$ for points in the dark region, and -1 for points in the light region.

2. Is $c_{\hat{\theta}}$ overfitting, underfitting or neither of the two? Briefly justify your choice. How would you change the current value of λ to improve performance?

2 pts $\square$

. . . .

. . . .

. . . .

3. Write down the general expression for the expected risk $R(\hat{c})$ of the classifier $\hat{c}$. Briefly explain how the test error is related to the expected risk.

2 pts $\square$

. . . .

. . . .

. . . .

Question 6: Feature Transformations (8 pts)

For each of the plots in Figure 3, provide the weights of a single layer perceptron that distinguishes the dark and the light regions (ignoring the boundary). If necessary, first apply a suitable transformation $\phi : \mathbb{R}^2 \to \mathbb{R}^n$, $(x_1, x_2) \mapsto \phi(x_1, x_2)$, such that the resulting regions are linearly separable.

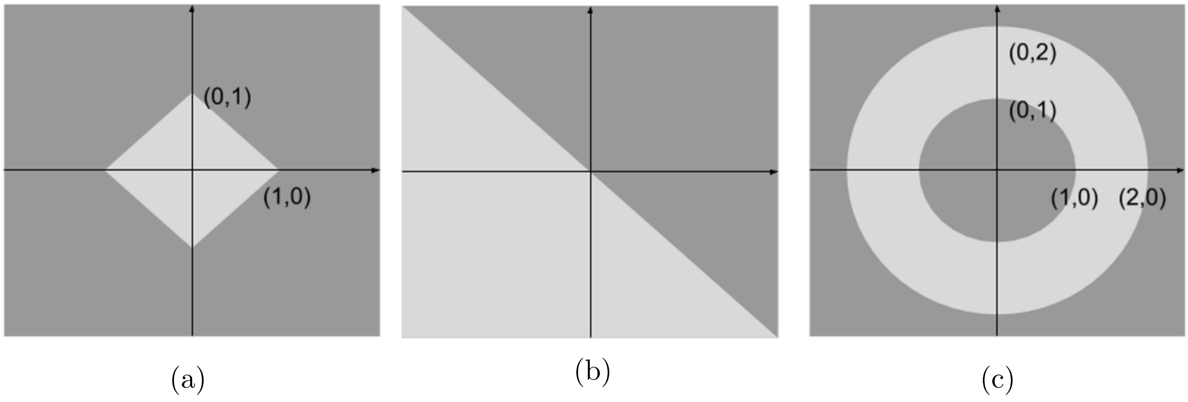


Figure 3

(a)

3 pts

. . . .

. . . .

. . . .

. . . .

(b)

2 pts

. . . .

. . . .

. . . .

. . . .

(c)

3 pts

. . . .

. . . .

. . . .

. . . .

Question 7: Properties of Perceptrons (5 pts)

Which of the following claims are true/false?

5 pts ☐

1) The solution of the perceptron algorithm can always be computed in closed form.
☐ True ☐ False

2) The solution of the perceptron algorithm does not depend on the learning rate value.
☐ True ☐ False

3) The perceptron can be trained in an online mode.
☐ True ☐ False

4) The perceptron algorithm converges for non-separable data.
☐ True ☐ False

5) The perceptron allows missclassifications.
☐ True ☐ False

Question 8: Information Criteria (4 pts)

1. What is the range of possible values that the Akaike Information Criterion (AIC) resp. the Bayesian Information Criterion (BIC) can take? (Provide an answer for each of them.)

2 pts ☐

....

....

....

2. Suppose we model a set of 67 data points with a distribution parameterized by 8 real numbers. Suppose the maximum likelihood of that dataset is $e^{-34.5}$. Calculate the AIC for this model (provide the final answer as a decimal number).

2 pts ☐

....

....

....

Question 9: Numerical Estimation (5 pts)

Which of the following claims are true/false?

5 pts ☐

1) Leave-One-Out cross validation is equivalent to n-fold cross validation, where n is the number of points in the dataset.
   ☐ True ☐ False

2) For each dataset of cardinality n there exists a model distribution with AIC not greater than $2n(1 + \log n)$.
   ☐ True ☐ False

3) The AIC is less or equal to the BIC for any model distribution and any dataset of cardinality at least 9.
   ☐ True ☐ False

4) The jackknife estimator is always consistent.
   ☐ True ☐ False

5) The jackknife estimator always reduces the absolute bias of its base estimator.
   ☐ True ☐ False

Question 10: Bagging (4 pts)

Suppose that you trained a random forest for classifying spam, which achieves very good performance on your training data, but very bad performance on your validation data. Mark all possible statements that could possibly explain this.

4 pts ☐

☐ The decision trees are too deep.

☐ You have too few decision trees in your ensemble.

☐ You are sampling your data points with replacement.

☐ When choosing your split, you are randomly sampling too many features.

Question 11: AdaBoost (6 pt)

Consider building an ensemble $f(x) = \text{sign}(\sum_{m=1}^{M} \alpha_m G_m)$ of decision stumps G_m with the AdaBoost algorithm (a decision stump is a binary classifier whose prediction depends only on the value of *one* coordinate). Figure 4 below shows several labeled points, as well as the first stump. The little arrow in the figure is the normal to the stump's decision boundary indicating the side where the stump predicts $+1$. All of the points have weight 1 at the beginning.

6 pts ☐

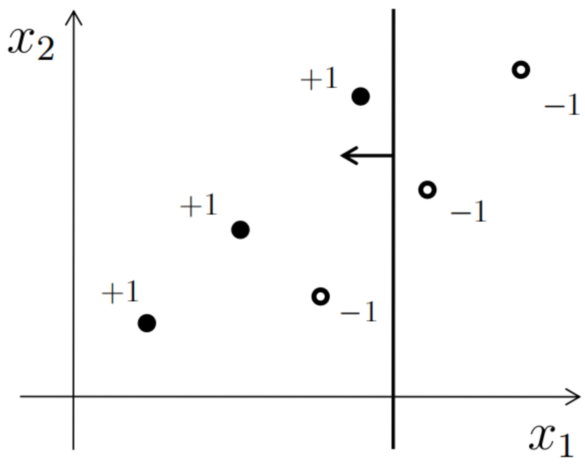


Figure 4: Labeled points and the first decision stump.

1. Circle all the point(s) in Figure 1 whose weight will be increased in the first iteration of AdaBoost.
2. In the same figure, draw the stump that AdaBoost selects at the next boosting iteration. Draw both the decision boundary and its orientation.
3. Consider the coefficients α_1, α_2 of the first and second stumps in the ensemble. Decide whether $\alpha_1 > \alpha_2$ or whether $\alpha_2 > \alpha_1$. Briefly justify your answer.

. . . .

. . . .

. . . .

Question 12: SVM and Kernels (13 pts)

1. Which of the following claims are true/false?

5 pts ☐

1) If a perceptron achieves zero training error on a dataset, then a hard margin SVM will also achieve zero training error on the same dataset.
☐ True ☐ False

2) If a hard margin SVM achieves zero training error on a dataset, then a soft margin SVM will also achieve zero training error on the same dataset, independent of the penalty strength.
☐ True ☐ False

3) Suppose that you have a linearly separable dichotomy and at least one element from each class. Then there exists a unique solution to the hard margin SVM problem.
☐ True ☐ False

4) If $k_1(\mathbf{x},\mathbf{y})$ and $k_2(\mathbf{x},\mathbf{y})$ are valid kernels then so is $k(\mathbf{x},\mathbf{y}) = k_1(\mathbf{x},\mathbf{y})k_2(\mathbf{x},\mathbf{y})$.
   ☐ True ☐ False

5) Every valid kernel $k(\mathbf{x},\mathbf{y})$, with $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$ can be represented with a finite dimensional feature transformation $k(\mathbf{x},\mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$, with $\phi : \mathbb{R}^d \to \mathbb{R}^n$ and $\langle \cdot, \cdot \rangle$ representing the standard Euclidean inner product.
   ☐ True ☐ False

2. Suppose you have a data set where one sample is d-dimensional $\mathbf{x}_i \in \mathbb{R}^d$, and you want to apply a feature transformation $\phi : \mathbb{R}^d \to \mathbb{R}^n$ that recovers the polynomial kernel $\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle = (c + \mathbf{x} \cdot \mathbf{y})^2$, where $\langle \cdot, \cdot \rangle$ represents the standard Euclidean inner product. What is the minimal required dimension n for such a feature transformation?

4 pts ☐

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

3. Suppose you have a data set $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}$ and after a feature transformation $\phi(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{c}$, with $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{c} \in \mathbb{R}^n$, a hard margin SVM is able to achieve zero training error. Show that the original data set was already linearly separable.

4 pts ☐

. . . .

. . . .

. . . .

. . . .

. . . .

Question 13: Structural SVMs **(11 pts)**

Structural SVMs (SSVMs) can be used to train classifiers that map elements $\mathbf{y}$ of an input space $\mathcal{Y}$ to elements z of a structured output space $\mathbb{K}$.

1. Write down the primal formulation of the soft-margin SSVM problem, denoting by $\mathbf{\Psi}(z, \mathbf{y})$ the joint features of inputs and outputs, and by $\Delta(z, z')$ the loss function.

2 pts ☐

. . . .

. . . .

. . . .

. . . .

. . . .

2. Assume that the weights $\mathbf{w}$ of an SSVM have been found by solving the above optimization problem. Write down the expression for the corresponding prediction function $h : \mathcal{Y} \to \mathbb{K}$.

2 pts ☐

. . . .

. . . .

3. Write down the expression for the Lagrangian of the primal formulation of the SSVM problem. To simplify the notation, you might write $\Psi_i(z) := \Psi(z_i, \mathbf{y}_i) - \Psi(z, \mathbf{y}_i)$ and $\Delta_i(z) := \Delta(z, z_i)$.

2 pts ☐

. . . .

. . . .

. . . .

. . . .

. . . .

points:

- -0.5 for missing $\sum_i \beta_i \xi_i$
- -0.5 for α not indexed by z

4. Derive the dual formulation of the SSVM problem.

5 pts

. . . .

Question 14: PAC learning (11 pts)

Consider the hypothesis class $\mathcal{C} := \left\{\mathbb{I}_{[\ell,\infty)} \mid \ell \in \mathbb{R}\right\}$ consisting of all indicator functions of intervals of the form $[\ell, \infty)$:

$$\mathbb{I}_{[\ell,\infty)}(x) = \begin{cases} 0 & x < \ell \\ 1 & x \geq \ell \end{cases}$$

Fix some $\ell^* \in \mathbb{R}$ and consider the classifier $c^* = \mathbb{I}_{[\ell^*,\infty)}$. Let $X_1, X_2, \ldots$ be i.i.d. random variables taking values in $\mathbb{R}$ and let $Y_i := c^*(X_i)$. Moreover, for $n \in \mathbb{N}$, let

$$X_{\min}^n := \min_{i \leq n, Y_i = 1} X_i \quad \text{and} \quad \hat{c}_n := \mathbb{I}_{[X_{\min}^n,\infty)}.$$

1. Let $\epsilon > 0$ and assume that there is a unique $\ell_\epsilon^+ \in \mathbb{R}$ such that $\mathbf{Pr}\left(\ell^* \leq X_i < \ell_\epsilon^+\right) = \epsilon$. Show that $\mathbf{Pr}\left(\ell_\epsilon^+ \leq X_{\min}^n\right) = (1-\epsilon)^n$.

3 pts ☐

. . . .

. . . .

. . . .

2. Show that if $n > \frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right)$, then $\mathbf{Pr}\left(\ell_\epsilon^+ \leq X_{\min}^n\right) < \delta$. *Hint:* $1 - z \leq \exp(-z)$.

3 pts ☐

. . . .

. . . .

. . . .

3. Show that $\mathcal{C}$ is efficiently PAC-learnable. *Hint: Prove* $\mathbf{Pr}\left(\mathcal{R}\left(\hat{c}_n\right) > \epsilon\right) = \mathbf{Pr}\left(\ell_\epsilon^+ \leq X_{\min}^n\right)$.

5 pts ☐

. . . .

. . . .

. . . .

Supplementary Sheet