ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Machine Learning
HS 2017

Prof. J. Buhmann

Final Exam

January 23th, 2018

First and Last name: _____

Student ID (Legi) Nr: _____

Signature: _____

General Remarks

- Please check that you have all 28 pages of this exam.
- There are 120 points, 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.
- Please use only a **black** or a **blue** pen to answer the questions.
- Provide only one solution to each exercise. Cancel invalid solutions clearly.
- If not indicated otherwise, multiple choice questions are graded as: 1 point per correct answer, -1 point per incorrect answer, non-negative total points in any case. Multiple correct answers are possible.

Grade: .....

	Topic	Max. Points	Points Achieved	Visum
1	Regression	6		
2	Elastic Net Regression	12		
3	Classification via density estimation	7		
4	Bayesian Learning	12		
5	Prior in Bayesian Inference	3		
6	MLE & Cramer-Rao bound	13		
7	Maximum Likelihood for Classification	5		
8	Validation	4		
9	k-fold Cross Validation	3		
10	Leave-One-Out Cross Validation	7		
11	k-Nearest Neighbours	8		
12	Linear Discriminant Functions	3		
13	Kernels	3		
14	SVM and Kernel Trick	4		
15	L2-SVM	14		
16	Neural Networks	7		
17	Ensemble methods	9		
Total		120		

Question 1: Regression: Warm-up (6 pts)

Which of the following claims are true/false?

6 pts

1) Although Lasso regression does not have a closed-form solution, its parameters can be learned through optimization methods like gradient descent.
   ☐ True ☐ False

2) The objective of Lasso regression is differentiable and convex.
   ☐ True ☐ False

3) Using basis expansions, regression models such as Ridge or Lasso regression can capture a non-linear relationship between the input $\mathbf{x}$ and the response y.
   ☐ True ☐ False

4) Let $\mathbf{x} \in \mathbb{R}^3$ and each entry of $\mathbf{x} = (x_1, x_2, x_3)$ be strictly greater than 0. For the model $y = \ln(x_1^{\beta_1} x_2^{\beta_2}) + \beta_3 x_3 + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$ is i.i.d. noise, the maximum likelihood parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ can be learned using linear regression.
   ☐ True ☐ False

5) In the bias-variance decomposition, the noise term can be eliminated by using a sufficiently complex model.
   ☐ True ☐ False

6) Among all linear estimators (biased and unbiased ones), least square regression always minimizes the expected mean squared error.
   ☐ True ☐ False

Question 2: Elastic Net Regression (12 pts)

The *elastic net* is a linear regression model that combines L_1 and L_2 regularization: Its objective function for given data $(\mathbf{X}, \mathbf{y}) \in \mathbb{R}^{n \times d} \times \mathbb{R}^n$ is defined by

$$L(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\rho\|\boldsymbol{\beta}\|_1 + \lambda(1-\rho)\|\boldsymbol{\beta}\|_2^2$$

for $\boldsymbol{\beta} \in \mathbb{R}^d$, where $\lambda > 0$ is the *regularization parameter* and $\rho \in [0, 1]$ is the L_1 *ratio*.

(*In the following multiple choice questions, 0.5 points are awarded per correct answer, -0.5 points per incorrect answer, non-negative total points in any case. Multiple correct answers are possible.*)

1. Why do we add the regularization terms? Mark the correct answer(s).

   2 pts

   ☐ Decreasing the estimator's variance.
   ☐ Decreasing the estimator's bias.
   ☐ Performing feature selection.
   ☐ Prevention of underfitting.

2. What is the purpose of the L_1 term? Mark the correct answer(s).

2 pts

- ☐ Performing feature selection.
- ☐ Compensating for overfitting.
- ☐ Making our model better interpretable.
- ☐ Smoothing.

3. When $\rho = 0$, what is the purpose of the L_2 term? Mark the correct answer(s).

2 pts

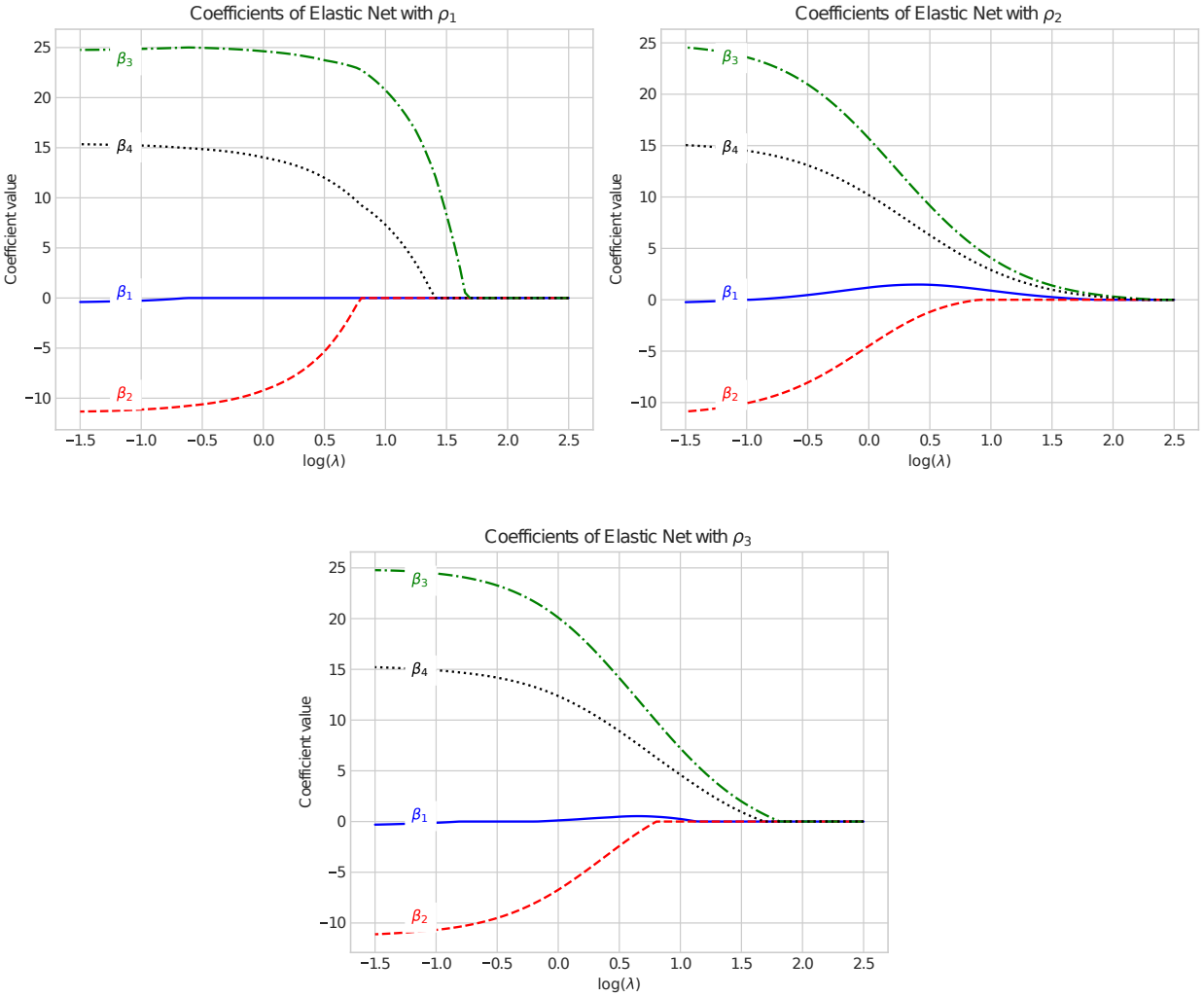
- ☐ Performing feature selection.
- ☐ Increasing the estimator's variance.
- ☐ Making the optimization problem strictly convex.
- ☐ Improving stability of the optimization procedure.

4. Derive the closed-form solution $\hat{\boldsymbol{\beta}}$ that minimizes the elastic net objective for the case $\rho = 0$.

4 pts

5. We now compare the behaviour of the elastic net for different values of the L_1 ratio ρ. Each plot in Figure 5 belongs to one of the values $\rho = 0.19, 0.69, 0.99$ and shows the results of different elastic net models that have been trained with respect to different values of λ. Which plot belongs to which value of ρ?

2 pts

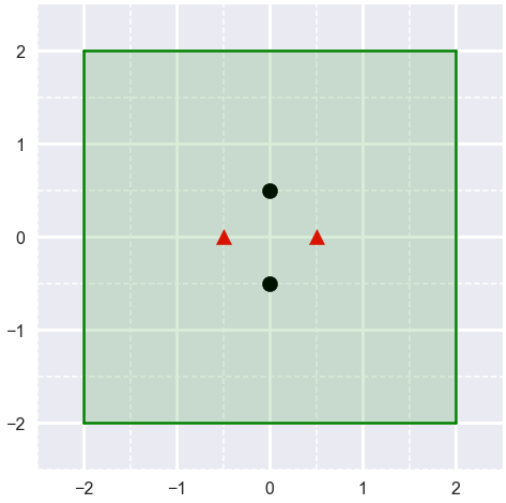


Question 3: Classification via Density Estimation (7 pts)

1. The following figure shows a dataset whose samples each belong to one of two possible classes $i \in \{0, 1\}$, represented by triangles and circles. In the following we consider classifiers based on density estimation. Recall that these work as follows: We first estimate the class densitites p_0, p_1 by $\hat{p}_0, \hat{p}_1$; then we classify new samples x according to the rule

$$f(\mathbf{x}) = \begin{cases} 0 & \text{if} \quad \hat{p}_0(\mathbf{x}) > \hat{p}_1(\mathbf{x}), \\ 1 & \text{if} \quad \hat{p}_0(\mathbf{x}) < \hat{p}_1(\mathbf{x}), \\ \text{N/A} & \text{otherwise.} \end{cases}$$

We assume that it is known that p_0, p_1 are zero outside of the shaded square.



(a) Assume that the class densities are estimated using a Parzen window density estimation with the window function

$$\phi(\mathbf{x}) = \begin{cases} 1/2 & \text{if} \quad \|\mathbf{x}\|_1 < 1, \\ 0 & \text{otherwise.} \end{cases}$$

where $\|\mathbf{x}\|_1 = |x_1| + |x_2|$; note that the window width is already chosen for you. Draw the region where no decision can be made (i.e., where "N/A" is returned) in the figure above.

4 pts ☐

2. Consider the generic density estimation formula discussed in class, $p(x) = \frac{K}{nV}$, where K is the number of points in the test region containing the data point x, V is the volume of the test region, and n is the total number of data points. Tick all true statements:

☐ Changing V corresponds to adjusting the window width in case of Parzen window density estimation.

☐ To counteract underfitting, one should decrease K.

☐ To counteract underfitting, one should increase V.

3 pts ☐

Question 4: Bayesian Learning (12 pts)

Assume that we are given a dataset $\mathcal{D}$ consisting of n random values $x_1, x_2, \ldots, x_n \in \mathbb{R}_{>0}$ drawn independently from the uniform distribution $\mathcal{U}(0, \theta)$ on an interval of the form $(0, \theta)$ for some unknown $\theta > 0$. Our aim is to find a good estimator for θ.

1. Compute the likelihood function $\theta \mapsto p(\mathcal{D}|\theta)$.

3 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

2. Assume now that $\mathcal{D} = \{x_1, x_2, x_3\}$ with $x_1 = 3$, $x_2 = 1.5$, $x_3 = 2$. Plot the likelihood function $\theta \mapsto p(\mathcal{D}|\theta)$.

2 pts ☐

3. Compute the maximum likelihood estimate $\hat{\theta}_{MLE}$ for θ given data $\mathcal{D} = \{x_1, \ldots, x_n\}$.

3 pts ☐

. . . . . .

. . . . . .

. . . . . .

4. The Pareto distribution Pareto(α, β) with parameters $\alpha, \beta > 0$ is the distribution on $\mathbb{R}$ whose probability density function is given by

$$p(\theta|\alpha, \beta) = \frac{\alpha\beta^\alpha}{\theta^{1+\alpha}}\mathbf{I}_{\{\theta \geq \beta\}}.$$

Show that if in the above situation we use Pareto(α, β) as the prior distribution for the parameter θ, the posterior distribution is Pareto(γ, δ) for certain parameters $\gamma, \delta > 0$. Compute γ and δ.

4 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

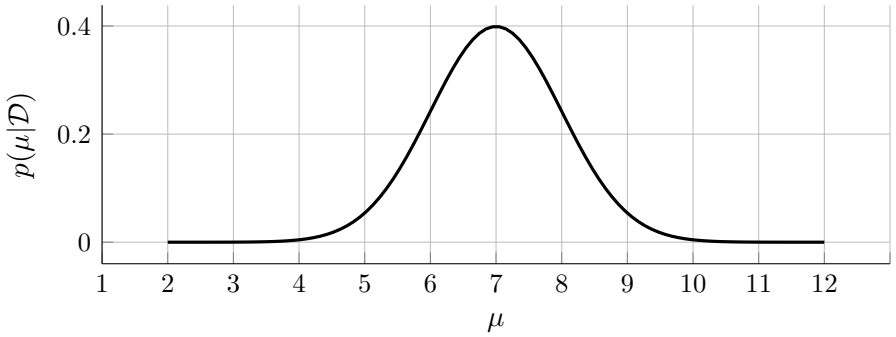
. . . . . .

. . . . . .

Question 5: Prior in Bayesian Inference (3 pts)

Suppose that we are given a dataset $\mathcal{D} = \{x_1, \ldots, x_n\}$ with samples $x_i \in \mathbb{R}$ drawn i.i.d. from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with unknown mean μ and variance σ^2.

1) If we assume a Gaussian prior $\mu \sim \mathcal{N}(0, \delta^2)$ for the mean, the density of the resulting posterior distribution of μ given $\mathcal{D}$ looks as follows:



Let now μ_{MAP} be the corresponding maximum a posteriori estimate, and let μ_{MLE} be the maximum likelihood estimate of μ given $\mathcal{D}$. Which of the following statements is true?

- ☐ $\mu_{\text{MAP}} = \mu_{\text{MLE}}$
- ☐ $\mu_{\text{MAP}} > \mu_{\text{MLE}}$
- ☐ $\mu_{\text{MAP}} < \mu_{\text{MLE}}$

1 pts ☐

2) We now change the prior to $\mu \sim \mathcal{N}(0, 10\delta^2)$. Mark all true statements.

- ☐ The mean of the posterior increases.
- ☐ The mean of the posterior decreases.
- ☐ The variance of the posterior increases.
- ☐ The variance of the posterior decreases.

2 pts ☐

Question 6: MLE & Cramer-Rao bound (13 pts)

Let's estimate the mean μ of a one-dimensional Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ from n samples $x_1, \ldots, x_n \in \mathbb{R}$ drawn independently at random.

1) Show that the maximum likelihood estimate of μ given $x_1, \ldots, x_n$ is

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

3 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

2) Show that $\hat{\mu}_n$ is an unbiased estimator of μ.

2 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

3) Indicate which of the following three terms is the *bias*, the *variance* resp. the *expected mean squared error* of the estimator $\hat{\mu}_n$.

1 pts ☐

$$\mathbf{E}_{\mathbf{x}}\left[(\widehat{\mu} - \mathbf{E}_{\mathbf{x}}[\widehat{\mu}])^2\right], \qquad \mathbf{E}_{\mathbf{x}}\left[(\mu - \widehat{\mu})^2\right], \qquad \mathbf{E}_{\mathbf{x}}[\widehat{\mu}] - \mu$$

4) Write down how the terms in 3) relate to each other.

2 pts

. . . . . .

. . . . . .

. . . . . .

. . . . . .

5) Denote by $p(\mathbf{x}|\mu, \sigma^2)$ the probability density of the random vector $\mathbf{x} = (x_1, \ldots, x_n)$. Recall that the *Fisher information* of the sample $\mathbf{x}$ about the parameter μ is defined by

$$I_n(\mu) = \mathbf{E}_{\mathbf{x}}[(\tfrac{\partial}{\partial \mu} \log p(\mathbf{x}|\mu, \sigma^2))^2].$$

Compute $I_n(\mu)$ as a function of n and σ^2.

3 pts

. . . . . .

. . . . . .

. . . . . .

. . . . . .

6) Use the Cramer-Rao bound to bound the expected mean squared error achieved by the estimator $\hat{\mu}_n$.

2 pts

. . . . . .

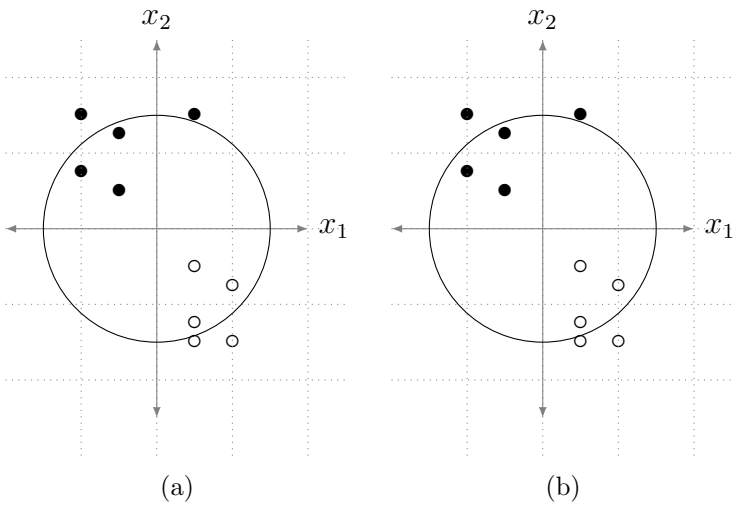
. . . . . .

. . . . . .

. . . . . .

Question 7: Maximum Likelihood for Classification (5 pts)

Assume that we are given a dataset $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{10}, y_{10})\}$ consisting of points $\mathbf{x}_i \in \mathbb{R}^2$ with labels $y_i \in \{-1, 1\}$, as depicted in the following figure, where the white points are those with $y_i = -1$, and the black points are those with $y_i = 1$:



(a) (b)

Figure 1: (Both figures are identical.)

Consider now the following generative model for the data:

$$P\left(y_i = y | \mathbf{x}_i, \mathbf{w}\right) = \begin{cases} 1 & \text{if } y(\mathbf{w}^\top \mathbf{x}_i) > 0, \\ 0 & \text{if } y(\mathbf{w}^\top \mathbf{x}_i) < 0, \end{cases}$$

where $\mathbf{w} = (w_1, w_2) \in \mathbb{R}^2$ is an unknown vector of length $\|\mathbf{w}\|_2 = 1$.

1. Determine the set of values of the likelihood function $\mathbf{w} \mapsto P(y_1, \ldots, y_{10} | \mathbf{x}_1, \ldots, \mathbf{x}_{10}, \mathbf{w})$.

1 pts ☐

. . . . . .

. . . . . .

2. Sketch the support of the likelihood function $\mathbf{w} \mapsto P(y_1, \ldots, y_{10} | \mathbf{x}_1, \ldots, \mathbf{x}_{10}, \mathbf{w})$ in Figure 1(a) (i.e., the region where its value is non-zero).

2 pts ☐

3. Add a black point $\mathbf{x}_{11}$, i.e. with $y_{11} = 1$, to Figure 1(b) such that the resulting likelihood function satisfies

$$\max_{\mathbf{w}} P(y_1, \ldots, y_{11} | \mathbf{x}_1, \ldots, \mathbf{x}_{11}, \mathbf{w}) = 0.$$

2 pts ☐

Question 8: Validation (4 pts)

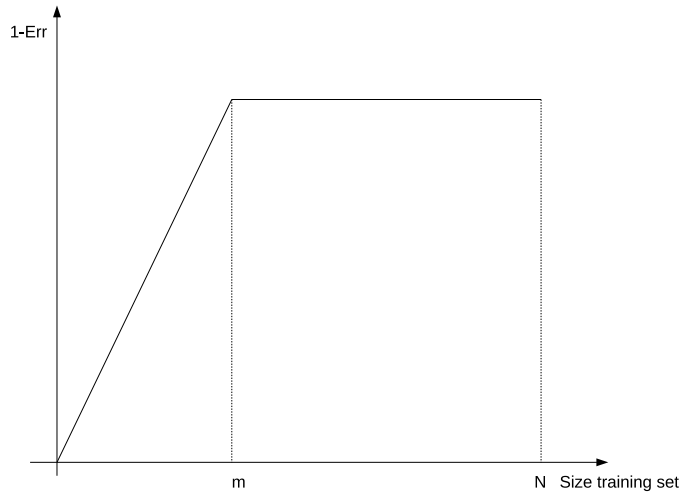
What is the goal of validation? Tick all correct answers. (*0.5 points per correct answer, -0.5 points per incorrect answer, non-negative total points in any case. Multiple correct answers are possible.*)

- ☐ Minimizing the training error.
- ☐ Performing model selection.
- ☐ Finding a minimum variance estimator.
- ☐ Finding a minimum bias estimator.
- ☐ Estimating model robustness.
- ☐ Estimating the generalization ability of the model.
- ☐ Choosing a good training and test set.
- ☐ Tuning hyper-parameters.

4 pts

Question 9: k-fold Cross Validation (3 pts)

Suppose that we have a dataset of size N and a learning algorithm for which the training set size and the expected generalization error Err are related as follows:



Here Err is the expected generalization error of a model obtained by training our model on a training set of a given size sampled from the dataset. Compute, for given $N > m$, the smallest k that results in an *unbiased* k-fold cross validation estimation of the generalization error.

3 pts

. . . . . .

. . . . . .

. . . . . .

Question 10: Leave-One-Out Cross Validation (7 pts)

Consider the two disjoint intervals $I_1 = [-2, -1]$ and $I_2 = [1, 2]$. Assume that we are given a training dataset $\{(x_i, y_i) \mid i = 1, \dots N\}$ consisting of i.i.d. data points generated as follows: First, the x_i are sampled with probability p from the uniform distribution on I_1, and with probability $1 - p$ from the uniform distribution on I_2; then, $y_i \in \{0, 1\}$ is sampled from a Bernoulli distribution with parameter u_1 if $x_i \in I_1$, and from a Bernoulli with parameter u_2 if $x_i \in I_2$:

$$y_i \sim \begin{cases} \mathrm{Ber}(u_1) & \text{if} \quad x_i \in I_1, \\ \mathrm{Ber}(u_2) & \text{if} \quad x_i \in I_2. \end{cases}$$

Assume now that we train a 1-nearest-neighbor (1NN) classifier on the training set.

1. Compute the expected training set error of the 1NN classifier. (Note: Training and testing happen on the same dataset.)

   1 pts ☐

   .....

   .....

2. Compute the expected leave-lone-out cross validation error of the 1NN classifier.

   3 pts ☐

   .....

   .....

   .....

   .....

   .....

   .....

3. Assume now that we get an independent test set that is generated in the same way as the training set, but with respect to different parameters $\bar{p}, \bar{u}_1, \bar{u}_2$. Compute the expected prediction error that the 1NN classifier incurs on the training set.

3 pts ☐

. . . . . .

. . . . . .

. . . . . .

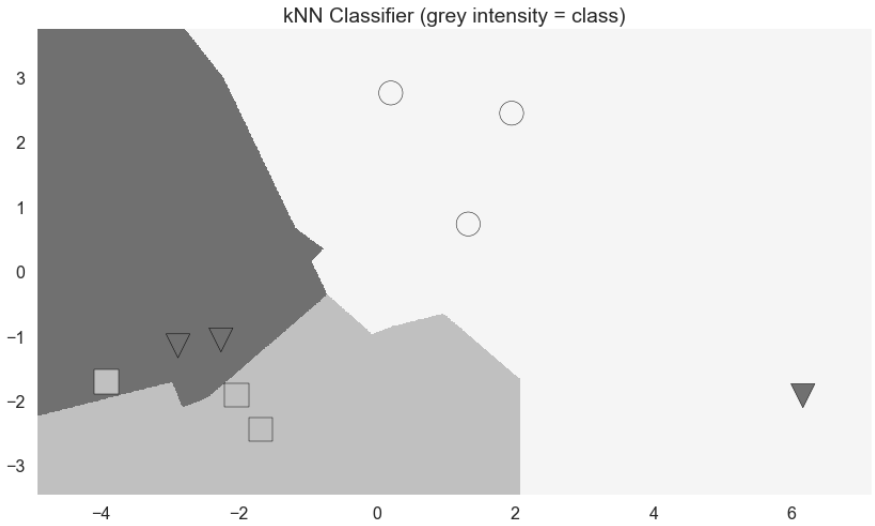
. . . . . .

. . . . . .

. . . . . .

Question 11: k-Nearest Neighbours (8 pts)

1. The figure below shows the output of a 3-class classification task obtained using the k-nearest neighbors algorithm for some value of k. Note that the symbols represent the training data.

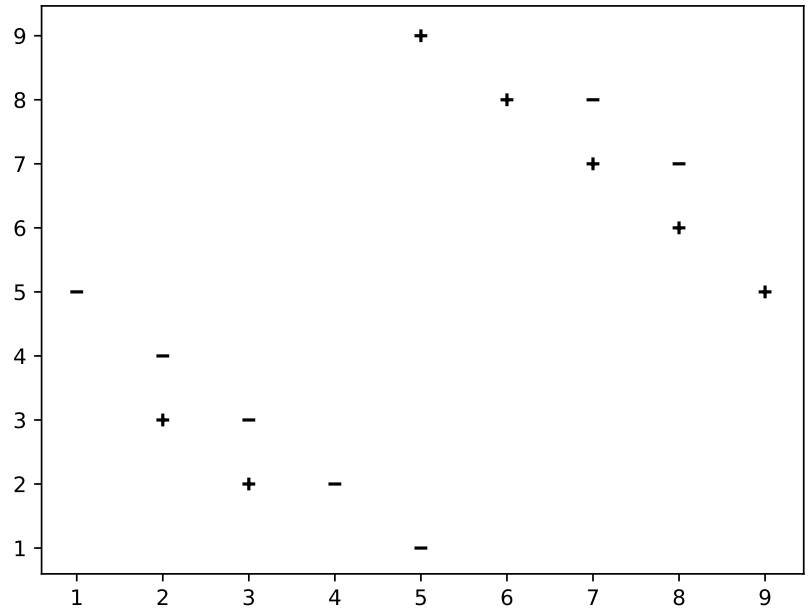


What value of k was used to obtain this classifier?

1 pts ☐

☐ $k = 1$.
☐ $k = 3$.
☐ $k = 5$.
☐ $k = 9$.

2. Consider a k-Nearest Neighbor classifier using the Euclidean distance metric on a binary classification task for the dataset depicted in the figure below.



(a) How would the point (2, 2) be classified using 1-Nearest Neighbor? **1 pts** ☐

. . . . . .

(b) Sketch the 1-Nearest Neighbor decision boundary in the figure. **3 pts** ☐

(c) Compute the leave-one-out cross-validation error that a 3-Nearest Neighbor classifier incurs when trained on this dataset. **3 pts** ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

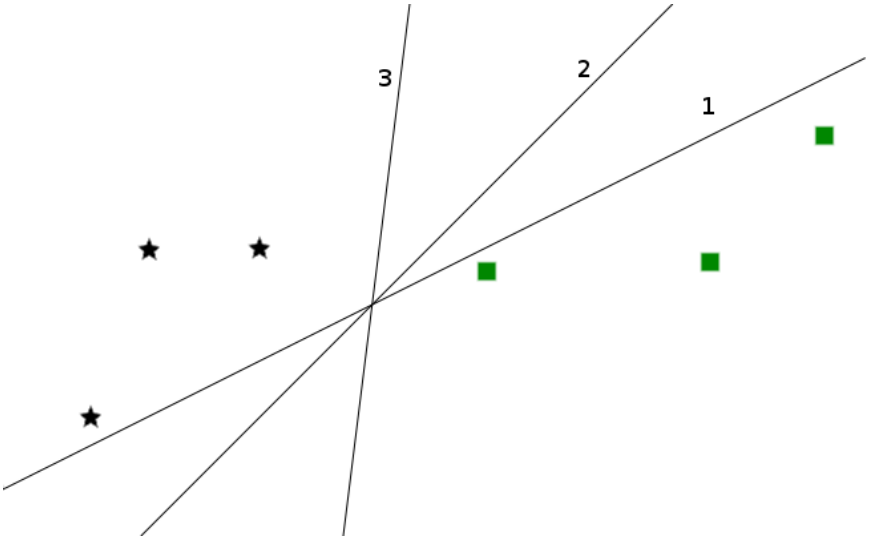
. . . . . .

. . . . . .

Question 12: Linear Discriminant Functions (3 pts)

1. Decide for each of the hyperplanes whether it was produced by running a Support Vector Machine (SVM), Fisher's Linear Discriminant Analysis (LDA), or a Perceptron on the dataset shown in the figure (each method was used exactly once).

2 pts



2. How many iterations does the Perceptron algorithm need to converge if we run it on the following dataset?

1 pts

- ☐ The answer depends on the initialization.
- ☐ The perceptron does not converge.
- ☐ 1.
- ☐ 3.

Question 13: Kernels (3 pts)

3 pts

1) If your input samples have two features, $x = (x_1, x_2)$, what is the dimension of the feature space for the kernel $K(x, x') = (x^T x')^d$ when $d = 2$?

   ☐ 3
   ☐ 5
   ☐ 6

2) If your input samples have three features, $x = (x_1, x_2, x_3)$, what is the dimension of the feature space for the kernel $K(x, x') = (x^T x' + c)^d$ when $d = 2$?

   ☐ 4
   ☐ 6
   ☐ 10

3) If you have 1000 input samples, each sample x has three features, $x = (x_1, x_2, x_3)$, and you choose to use the kernel $K(x, x') = (x^T x' + c)^d$ with $d = 2$, how many entries does the corresponding Gram matrix have?

   ☐ 1,000,000
   ☐ 6,000,000
   ☐ 10,000,000

Question 14: SVM and Kernel Trick (4 pts)

Which of the following claims are true/false?

4 pts

1. Every kernel corresponds to some transformation $\phi(x)$ which converts x to a strictly finite dimensional space.
   ☐ True ☐ False

2. The SVM classifier in its original form provides us with probability estimates.
   ☐ True ☐ False

3. We know that support vectors lie on the boundary and between margins, but we cannot explicitly identify them.
   ☐ True ☐ False

4. A standard way to counteract overfitting of a SVM is to use a more complex kernel (e.g., a polynomial kernel of degree 3 instead of degree 2).
   ☐ True ☐ False

Question 15: L_2-SVM (14 pts)

Assume that we modify the standard soft margin SVM by using an L_2 penalty on the slack variables (instead of the usual L_1 penalty). In other words, given a training set consisting of N labeled points $(\mathbf{x}_i, y_i)$ with $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \{-1, +1\}$, we consider the constrained optimization problem for the vector $\tilde{\mathbf{w}} = (\mathbf{w}, w_0) \in \mathbb{R}^{D+1}$ describing the separating hyperplane whose primal formulation is

$$\begin{aligned} \underset{\tilde{\mathbf{w}},\xi}{\text{minimize}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{1}{2}C\sum_{i=1}^{N}{\xi_i}^2 \\ \text{subject to} \quad & y_i(\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_\mathbf{i}) \geq 1 - \xi_i, \qquad \forall i, \\ & \xi_i \geq 0 \qquad\qquad\qquad \forall i. \end{aligned}$$

Here the ξ_i denote slack variables, and $\tilde{\mathbf{x}}_i := (\mathbf{x}_i, 1) \in \mathbb{R}^{D+1}$.

1. In the following, you are asked to show step by step that deleting the constraints $\xi_i \geq 0$ does not affect the solution of the optimization problem.

   6 pts ☐

   - Step 1: Assume that there exists a solution $(\tilde{\mathbf{w}}, \xi)$ so that ξ has at least one component $\xi_i < 0$. Is there another solution with $\xi_i \geq 0$ that leads to the same objective?

     . . . . . .

     . . . . . .

     . . . . . .

   - Step 2: Does the optimization problem have a special property that allows you to conclude that it attains a unique solution?

     . . . . . .

     . . . . . .

     . . . . . .

   - Step 3: Can you conclude that each solution of the optimization problem without the constraints $\xi_i \geq 0$ fulfills the constraint $\xi_i \geq 0$ automatically?

     . . . . . .

     . . . . . .

     . . . . . .

     . . . . . .

     . . . . . .

2. Write down the Lagrangian for the problem **2 pts** ☐

$$\begin{aligned} \underset{\tilde{\mathbf{w}},\xi}{\text{minimize}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{1}{2}C\sum_{i=1}^{N}{\xi_i}^2 \\ \text{subject to} \quad & y_i(\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_\mathbf{i}) \geq 1 - \xi_i, \qquad \forall i. \end{aligned}$$

. . . . . .

. . . . . .

. . . . . .

. . . . . .

3. Derive the dual formulation of the problem. **5 pts** ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

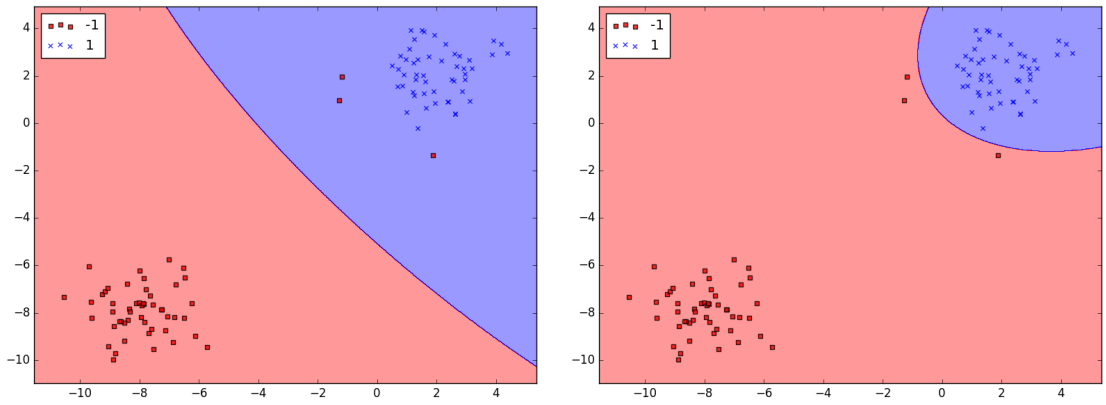
. . . . . .

. . . . . .

. . . . . .

4. The following plots show the decisions boundaries produced by running a kernel version of the L_2-SVM on a dataset, once with $C = 1$, and once with $C = 1000$. Decide which plot belongs to which value of C.

1 pts ☐



Question 16: Neural Networks (7 pts)

Given N input/output pairs $(\mathbf{x}_n, \mathbf{y}_n)$, where $\mathbf{x}_n \in \mathbb{R}^I$ and $\mathbf{y}_n \in \mathbb{R}^D$ $\forall n \in \{1, \ldots, N\}$, we want to train a neural network to predict $\mathbf{y}$ for a given $\mathbf{x}$. Given the input $\mathbf{x}_n$, the output of the neural network is denoted as $\hat{\mathbf{y}}_n = \hat{\mathbf{y}}(\mathbf{W}, \mathbf{x}_n)$ and the weight parameters as entries of the matrix $\mathbf{W}$.

1. Why are the activation functions in the hidden layers generally non-linear?

1 pts ☐

. . . . . .

. . . . . .

2. Name three regularization techniques and briefly describe each technique.

3 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

3. Which of the following claims are true/false? **3 pts**

(a) The more complex a neural network is, the better its generalization capacity is.
☐ True ☐ False

(b) The XOR task can be learned by a neural network consisting of two layers, i.e. one input layer and one output layer.
☐ True ☐ False

(c) Every continuous function can be approximated by a fully connected feedforward neural network with 4 hidden layers and sigmoid activation functions.
☐ True ☐ False

Question 17: Ensemble methods (9 pts)

1. Which of the following claims are correct? **3 pts**

(a) A decision tree is equivalent to a random forest consisting of one tree.
☐ True ☐ False

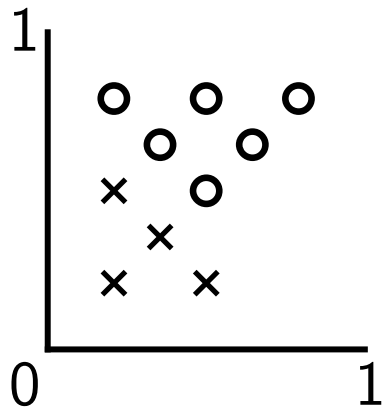
(b) Given an arbitrary collection of base classifiers and a classification performance target, an ensemble can be constructed from that collection that achieves the target.
☐ True ☐ False

(c) Suppose you are given one base classification model whose parameters are trained by a deterministic algorithm. Assume that the training strategy for a bagging ensemble was changed to remove the bootstrapping process. Thus, each base classifier is trained on the original training set. How would the classification performance change for increasing ensemble size?

☐ Decrease.
☐ Increase.
☐ Stay the same.
☐ Not clear given the information.

2. What is the smallest number of axis-oriented, linear decision boundaries needed to partition the data given below? Draw one possible solution on the grid.

2 pts ☐



3. What is the smallest number $b_{\min}$ of linear decision boundaries (not necessarily axis-oriented) needed to perfectly partition the data? Write down the equation(s) describing the line(s).

2 pts ☐

. . . . . .

. . . . . .

. . . . . .

4. Write down a linear transformation of the data so that it can be perfectly partitioned by $b_{\min}$ axis-oriented decision boundaries.

2 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .