ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

# Advanced Machine Learning

Fall 2019

*Prof. Joachim M. Buhmann*

## Final Exam

January 21, 2020

First and last name: _____

Student ID number: _____

Signature: _____

# General Remarks

- Please check that you have all 22 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. The maximal number of points is not required for the top grade 6.0.
- 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. Pencils are not allowed.
- Provide only one solution to each exercise. Invalid solutions have to be cancelled 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.

| | Topic | Points | Points achieved | Checked |
|---|---|---|---|---|
| 1 | Linear regression | 4 | | |
| 2 | LASSO as MAP | 8 | | |
| 3 | Bootstrapping | 6 | | |
| 4 | Bayesian modeling | 5 | | |
| 5 | MAP and MLE | 5 | | |
| 6 | Multiclass classification | 4 | | |
| 7 | Linear discriminators | 11 | | |
| 8 | Gaussian processes | 12 | | |
| 9 | Kernels | 7 | | |
| 10 | SVMs | 12 | | |
| 11 | Bagging | 8 | | |
| 12 | Clustering | 12 | | |
| 13 | PAC learning | 6 | | |
| Total | | 100 | | |

## Question 1: Linear regression (4 pts)

Which of the following claims are true/false? *(1 point per correct answer, -1 point per incorrect answer, non-negative total points in any case)*

1. Let $\mathbf{X}$ be a centered $n \times d$ matrix whose rows are the input vectors $\mathbf{x}_i \in \mathbb{R}^d$ of a training set $\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1, \ldots, n\} \subset \mathbb{R}^d \times \mathbb{R}$, and let $\mathbf{y} \in \mathbb{R}^n = (y_1, \ldots, y_n)$.

**2 pts** ☐

   a) If a closed form linear regression solution $\hat{f}(\mathbf{x}) = \mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ exists, then $\hat{f}(\mathbf{x})$ is the projection of $\mathbf{x}$ onto a hyperplane that intersects the origin.

   ☐ True ☐ False

   b) Linear ridge regression does not have a closed form solution if $n < d$.

   ☐ True ☐ False

2. Consider now the regularized least squares optimization problem given by

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \big(y_i - \sum_{j=1}^{d} x_{i,j} \beta_j\big)^2 + \lambda \sum_{j=1}^{d} |\beta_j|^q \right\}.$$

**2 pts** ☐

   a) This optimization problem is convex for every $q > 0$.

   ☐ True ☐ False

   b) This optimization problem is equivalent to minimizing the unregularized least squares error $\sum_{i=1}^{n} \big(y_i - \sum_{j=1}^{d} x_{i,j} \beta_j\big)^2$ subject to the constraint $\sum_{j=1}^{d} |\beta_j|^q \leq \mu$ for an appropriate value of the parameter $\mu$.

   ☐ True ☐ False

## Question 2: LASSO regression as MAP estimation (8 pts)

Consider a data set $\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1, \ldots, n\} \subset \mathbb{R}^d \times \mathbb{R}$. Assume that the data distribution from which $\mathcal{D}$ is sampled has the property that for every $\mathbf{x} \in \mathbb{R}^d$, the conditional distribution of the target is Gaussian, i.e.,

$$Y|(X = \mathbf{x}, \beta) \sim \mathcal{N}\left(\mathbf{x}^\top \beta, \sigma^2\right),$$

for some known fixed $\sigma > 0$. Here $Y$ and $X$ denote the random variables of which the $\mathbf{x}_i$, $y_i$ are realizations.

Recall that performing LASSO regression means solving the optimization problem

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta^\top \mathbf{x}_i\right)^2 + \lambda \sum_{j=1}^{d} |\beta_j| \right\}. \tag{1}$$

In this problem, the goal is to show that LASSO regression is equivalent to finding the maximum a posteriori (MAP) estimate in the Bayesian setting in which we assume that the weights $\beta_j$ have i.i.d. Laplacian prior distributions:

$$p(\beta_j) = \frac{1}{2b} \exp(-\frac{|\beta_j|}{b}) \qquad \text{for all } j \in \{1, 2, ..., d\}$$

**Hint:** For simplicity, you may assume that the data set $\mathcal{D}$ is *centered* in the following.

1. Write down the natural logarithm of the *un-normalized* posterior distribution of $\beta$. Do not stop at a symbolic formulation, but actually plug in the explicit probability density functions for the data points. "Un-normalized" means that you do not have to worry about the evidence, but only compute a term that is proportional to the posterior.

   **4 pts** ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

2. Show that maximization with respect to $\beta$ of the term we derived in a) for a given value of $b$ is equivalent to minimization of the LASSO objective in equation $(1)$ for a value $\lambda \equiv \lambda(b)$. Derive the formula describing how $\lambda(b)$ depends on $b$.

**4 pts** ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

## Question 3: Bootstrapping (6 pts)

Recall that the naive bootstrapping estimate of the error of a classifier is defined as

$$\widehat{Err}_{boot} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N} L(y_i, \hat{c}^{*b}(x_i)),$$

where $B$ is the number of training sets (each of the same size $N$ as the original data set), and $\hat{c}^{*b}(x_i)$ is the value at $x_i$ predicted by the model fitted to the $b$th bootstrap dataset. Here $L$ is the 0/1-loss, i.e., $L(y, c(x)) = 1_{\{c(x)\neq y\}}$.

Consider now a 1-nearest neighbor classifier applied to a two-class classification problem with the same number of observations in each class, in which the features and class labels are *independent*.

a) Compute the true error rate of the 1-nearest neighbor classifier.

**2 pts** ☐

. . . . . .

. . . . . .

b) Give an expression in terms of $N$ for the error estimated by $\widehat{Err}_{boot}$ as $N \to \infty$.

**Hint:** For a data point $x_i$, compute the probability that $x_i$ is contained in a given bootstrap data set.

**2 pts** ☐

. . . . . .

. . . . . .

. . . . . .

c) Give a reason why $\widehat{Err}_{boot}$ does *not* provide a good estimate for the true prediction error.

**2 pts** ☐

. . . . . .

. . . . . .

. . . . . .

## Question 4: Bayesian modeling (5 pts)

Consider a Bayesian model with data $y$ and parameters $\beta$. Using the terms $p(y)$, $p(\beta)$, $p(y \mid \beta)$, and $p(\beta \mid y)$, ...

a) ... write the terms for the likelihood, prior, posterior and evidence.

**1 pts** ☐

. . . . . .

. . . . . .

b) ... write the expression for the evidence in terms of the likelihood and prior.

**1 pts** ☐

. . . . . .

. . . . . .

c) . . . write the expression for the posterior in terms of the likelihood, prior and evidence.

**1 pts**

. . . . . .

. . . . . .

d) . . . write the expression for $p(y^* \mid y)$ needed for predicting the values of unseen data $y^*$ given the observed data $y$.

**2 pts**

. . . . . .

. . . . . .

## Question 5: MAP and MLE (5 pts)

The Poisson distribution is a commonly used distribution for modelling count data. Its density is given by

$$\text{Poisson}(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \tag{2}$$

where $x \in \mathbf{N}_0$ and $\lambda \in \mathbf{R}^+$. Assume $x_1, \ldots, x_n$ are i.i.d. samples from a Poisson distribution with parameter $\lambda$.

a) Write the log likelihood of $x_1, \ldots, x_n$.

**2 pts**

. . . . . .

. . . . . .

. . . . . .

b) Compute the maximum likelihood estimate of $\lambda$.

**3 pts**

. . . . . .

. . . . . .

. . . . . .

## Question 6: Multiclass classification (4 pts)

Suppose that you are given a classifier with the following confusion matrix.

| | | Actual | | |
|---|---|---|---|---|
| | | Class 1 | Class 2 | Class 3 |
| Predicted | Class 1 | 1 | 10 | 16 |
| | Class 2 | 12 | 42 | 1 |
| | Class 3 | 7 | 8 | 3 |

a) What is the accuracy of this classifier? Please write your answer using two decimal places, e.g., 0.23.

**2 pts** ☐

.....

.....

b) What is the balanced accuracy of this classifier? Balanced accuracy is defined as the average of recall obtained on each class, and recall for a given class is $TP/(TP + FN)$ with $TP$ = true positives, $FN$ = false negatives. Please write your answer using two decimal places, e.g., 0.23.

**2 pts** ☐

.....

.....

## Question 7: Linear discriminators (11 pts)

1) Write down the formula $\mathcal{L}(y, z)$ for each of the loss functions listed below, where $y \in \mathbb{R}$ is the prediction output by a discriminant function and $z \in \{\pm 1\}$ is the true label.

**3 pts** ☐

a) Hinge loss: $\mathcal{L}(y, z) =$.....

b) Exponential loss: $\mathcal{L}(y, z) =$.....

c) Logistic loss: $\mathcal{L}(y, z) =$.....

2) Which of these loss functions is *not* everywhere differentiable? *(There may be more than one correct answer. 1 point per correct answer, -1 point per incorrect answer, non-negative total points in any case)*

**1 pts**

- ☐ Hinge loss
- ☐ Exponential loss
- ☐ Logistic loss

3) Assume $y = \mathbf{w}^T\mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^k$ is the feature vector and $\mathbf{w} \in \mathbb{R}^k$ is the weight vector, and the *exponential loss* function $\mathcal{L}(y, z)$ is used. Compute the update rule for $\mathbf{w}$ using Newton's method. (Write down the necessary steps for deriving your solution.)

**5 pts**

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

4) Design in the following framed area a dataset of 2 classes with 3 samples in each class such that a perceptron will never converge with the perceptron algorithm. (Use ▲ to represent samples in class 1 and • to represent samples in class 2.)

**2 pts**

## Question 8: Gaussian processes (12 pts)

1. Let $\mathcal{D} = \{(\boldsymbol{x}_i, f_i) \mid i = 1, \ldots, N\}$ denote a set of observed data, where $f_i = f(\boldsymbol{x}_i)$ is the noise-free observation of the function evaluated at $\boldsymbol{x}_i$. We write $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^T \in \mathbb{R}^{N \times d}$ for the matrix containing the $\boldsymbol{x_i} \in \mathbb{R}^d$ as rows.

   Given a new test set $\mathbf{X}_*$, we want to predict the corresponding function outputs $\boldsymbol{f}_*$. In a Gaussian process model, the joint distribution of $\mathbf{f}$ and $\mathbf{f}^*$ has the form

$$\begin{pmatrix} \mathbf{f} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N}\left( \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}_* \end{pmatrix}, \begin{pmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{pmatrix} \right),$$

   where $\mathbf{K} = \kappa(\mathbf{X}, \mathbf{X})$, $\mathbf{K}_* = \kappa(\mathbf{X}, \mathbf{X}_*)$, $\mathbf{K}_{**} = \kappa(\mathbf{X}_*, \mathbf{X}_*)$ and $\kappa(\cdot)$ is a kernel function.

   a) Derive the predictive distribution $p(\mathbf{f}_* | \mathbf{X}_*, \mathbf{X}, \mathbf{f})$. Hint: Use the rules for conditioning Gaussian distributions.

   **2 pts** ☐

   . . . . . .

   . . . . . .

   . . . . . .

   . . . . . .

   b) Often, the observation are noisy, i.e., $y_i = f_i + \epsilon_i$, with i.i.d. noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Show that the covariance matrix has now the form $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} + \sigma^2 \boldsymbol{I}$, where $\boldsymbol{\Sigma}$ denotes the covariance matrix of the noise-free latent function.

   **4 pts** ☐

   . . . . . .

   . . . . . .

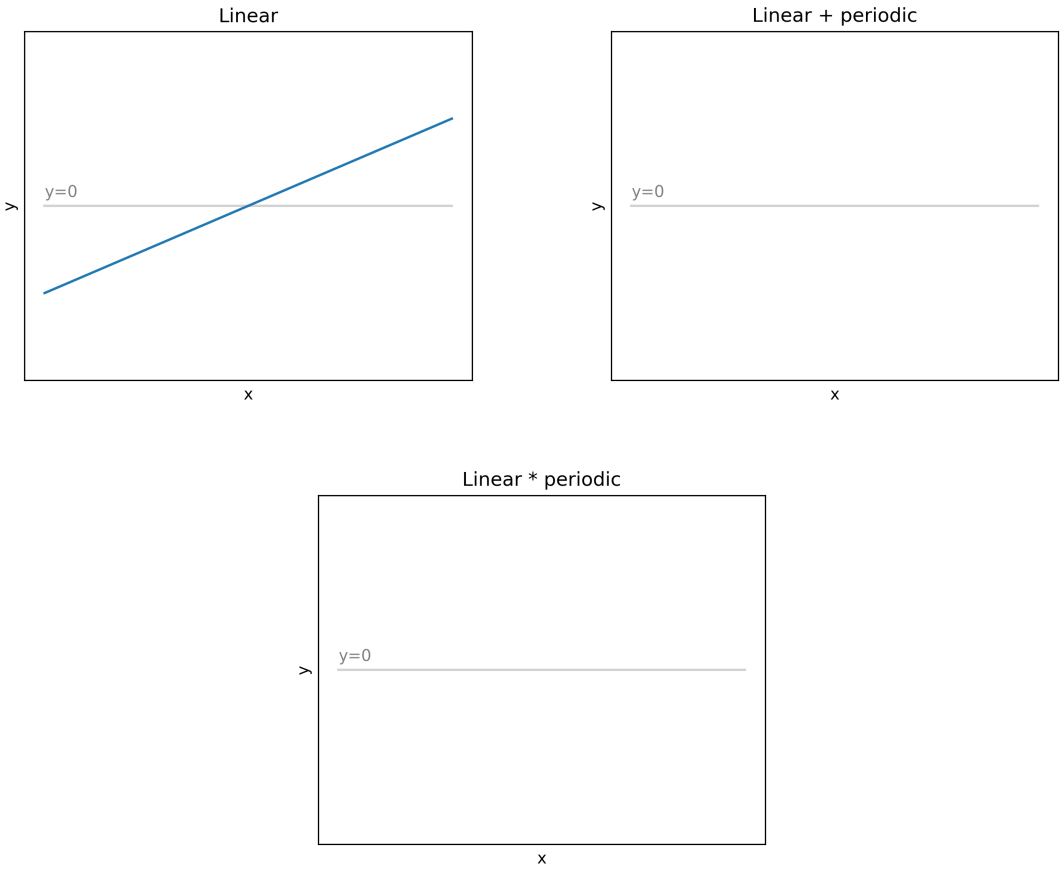
   . . . . . .

   . . . . . .

   c) To gain more flexibility in modelling or to impose certain constraints, one can combine kernel functions. Sketch how a typical function drawn from a Gaussian process looks like (similarly to the sketch for a linear kernel below) if . . .
      - . . . an additive combination of linear and periodic kernels is used.
      - . . . a multiplicative combination of linear and periodic kernels is used.

      In your sketch, focus on the most important features.

   **2 pts** ☐




2. Decide which of the given alternatives is correct. *(1 point per correct answer, -1 point per incorrect answer, non-negative total points in any case)*

**4 pts**

1) Consider the RBF kernel. As we substantially decrease the length scale, keeping the data and other parameters fixed, the more "wiggly" the resulting curves will be.
☐ True ☐ False

2) In a GP with an RBF kernel, the covariance of two points depends on their relative position, whereas in a GP with a linear kernel, the covariance depends also on their absolute location.
☐ True ☐ False

3) Conditioning in Gaussian Processes corresponds to integrating over one of the dimensions.
☐ True ☐ False

4) In conventional regression methods we typically allow for variation in both, the model class and coefficients, whereas in Gaussian Processes regression the model class is fixed.
☐ True ☐ False

## Question 9: Kernels (7 pts)

1. In the following problem, your task is to decide whether the given matrices are valid to be used as kernel matrices for a SVM. *Provide a brief justification in every case.*

   a) Is $\begin{bmatrix} 2 & 2 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$ a valid kernel matrix? **1 pts** ☐

   .....

   .....

   .....

   b) Is $\begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix}$ a valid kernel matrix? **2 pts** ☐

   .....

   .....

   .....

   c) Is $\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -0.5 \\ 0 & -0.5 & 3 \end{bmatrix}$ a valid kernel matrix? **2 pts** ☐

   .....

   .....

   .....

2. Prove or disprove: A kernel matrix can have a negative element in the diagonal.

   **2 pts** ☐

   .....

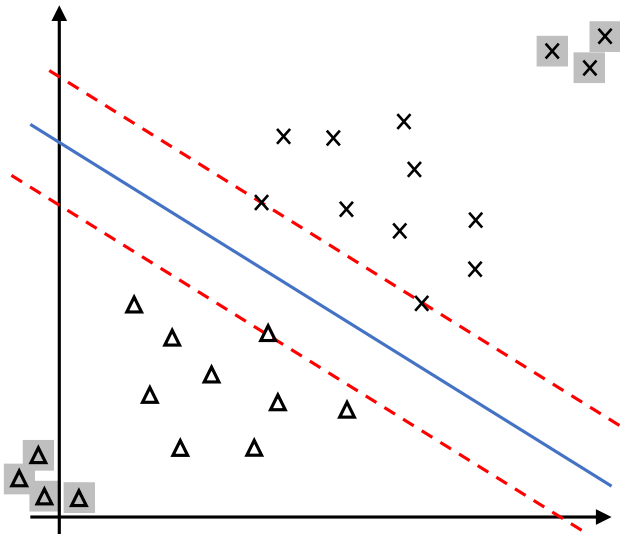
   .....

   .....

## Question 10: SVMs (12 pts)

1. Data points are depicted by crosses and triangles for the two classes, outliers are color marked with a gray background. Identify the support vectors in the following figure.

**1 pts** ☐



2. Which of the following claims are true/false? *(1 point per correct answer, -1 point per incorrect answer, non-negative total points in any case)*

**5 pts** ☐

(a) If you remove all the support vectors from the data in the above figure, the decision boundary will change.

☐ True ☐ False

(b) If you remove all outlier points in the above figure, the decision boundary will change.

☐ True ☐ False

(c) In soft-margin SVM with $C = \infty$ and linearly separable data, any hyperplane that completely separates the data is optimal. ($C$ is the weight of the slack in the loss.)

☐ True ☐ False

(d) In soft-margin SVM with $C$ close to zero, the optimal hyperplane maximizes the margin between most points but may misclassify some of them.

☐ True ☐ False

(e) In soft-margin SVM, we need to compute the slack variables for the test points in order to make a prediction.

☐ True ☐ False

3. Consider the following modified soft-margin SVM loss:

$$\min_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\|\xi\|^2 \tag{3}$$

$$\begin{aligned} \text{subject to:} \quad y_i \mathbf{w}^T \mathbf{x}_i &\geqslant 1 - \xi_i \quad \text{for all } i = 1, \ldots, n \\ \xi_i &\geqslant 0 \quad \text{for all } i = 1, \ldots, n. \end{aligned}$$

where we penalize the $L^2$-norm of the slack $\xi$ rather than the $L^1$-norm.

a) Compute the Lagrangian of the loss in $(3)$. **2 pts** ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

b) Write down the stationary conditions of the Lagrangian with respect to the primal variables. **2 pts** ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

c) Write down the dual problem for the primal problem given in $(3)$. **2 pts** ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

## Question 11: Bagging (8 pts)

1) Write down the variance of the random forest predictor $\hat{f}_B(x) = \frac{1}{B}\sum_{i=1}^{B} T_i(x)$, where each $T_i$ is a decision tree. You can use $\sigma^2$ to represent the variance of each individual $T_i$ and $\rho \equiv \rho(T_i, T_j)$ to represent the pairwise correlation between different decision trees (in particular, you may assume that this value is the same for all pairs). Recall that

$$Var(\sum_{i=1}^{N} X_i) = \sum_{i=1}^{N}\sum_{j=1}^{N} Cov(X_i, X_j) \quad \text{and} \quad Cov(X_i, X_j) = \frac{\rho(X_i, X_j)}{\sqrt{Var(X_i)Var(X_j)}}$$

**4 pts** ☐

.....

.....

.....

.....

.....

.....

.....

.....

2) What is the ideal value of $\rho$ and why? Choose **one** of the following alternatives:

**1 pts** ☐

☐ Small, since the aim of bagging is to reduce variance of the aggregate predictor $\hat{f}_B(x)$.

☐ Small, since the aim of bagging is to reduce variance of each base predictor $T_i(x)$.

☐ Large, since the aim of bagging is to reduce variance of the aggregate predictor $\hat{f}_B(x)$.

☐ Large, since the aim of bagging is to reduce variance of each base predictor $T_i(x)$.

☐ Does not matter, the predictor will have low variance in any case.

3) For each of the following situations, mark whether we would expect an increase or decrease in $\rho$:

**3 pts** ☐

(a) The number of predictor variables used at each split point is increased.

☐ $\rho$ increases.

☐ $\rho$ decreases.

(b) Duplicate predictor variables are added to the data set.

☐ $\rho$ increases.

☐ $\rho$ decreases.

(c) The bootstrap sample size is decreased from $N$ to $\sqrt{N}$, where $N$ is the total data set size.

☐ $\rho$ increases.

☐ $\rho$ decreases.

## Question 12: Clustering (12 pts)

1. Let us first look at the space of all clusterings and its size.

   a) What is the number of different ways to cluster $n$ data points into two non-empty disjoint clusters?

   **2 pts** ☐

   . . . . . .

   . . . . . .

   b) What is the number of different ways to cluster $n$ data points into $k$ or less disjoint clusters?

   **2 pts** ☐

   . . . . . .

   . . . . . .

2. Consider two clustering solutions $\mathcal{U} = \{U_1, \ldots, U_R\}$ and $\mathcal{V} = \{V_1, \ldots, V_R\}$ obtained from two different methods that cluster a dataset $X = \{x_1, \ldots, x_N\}$ into $R$ clusters. Recall that the *purity* is defined as

$$\text{purity}(\mathcal{U}, \mathcal{V}) = \frac{1}{|X|} \sum_{V \in \mathcal{V}} \max_{U \in \mathcal{U}} |U \cap V|,$$

   where $|\cdot|$ denotes cardinality.

a) Prove the following inequality:

$$\text{purity}(\mathcal{U}, \mathcal{V}) \geq \frac{\max_i |U_i|}{N}.$$

**2 pts**

. . . . . .

. . . . . .

b) Assume that each cluster in $\mathcal{U}$ has at least $k$ objects that belong to the same cluster in $\mathcal{V}$. What is the smallest possible purity of clusterings $\mathcal{U}$ and $\mathcal{V}$?

**2 pts**

. . . . . .

. . . . . .

3. Write the objective optimized by the $k$-means algorithm. Briefly explain your notation.

**2 pts**

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

4. Which of the following claims are true/false? *(1 point per correct answer, -1 point per incorrect answer, non-negative total points in any case)*

**2 pts**

1) In the Chinese restaurant process, new customers prefer to join big groups rather than create new ones.

☐ True ☐ False

2) For larger values of the concentration parameter $\alpha$, the Dirichlet process produces fewer distinct clusters.

☐ True ☐ False

## Question 13: PAC Learning (6 pts)

1. Which of the following claims are true/false? *(1 point per correct answer, -1 point per incorrect answer, non-negative total points in any case)*

**3 pts**

Let $\mathcal{C}$ and $\mathcal{C}'$ be concept classes and let $\mathcal{H}$ and $\mathcal{H}'$ be hypothesis classes.

a) If $\mathcal{C}$ is PAC-learnable from $\mathcal{H}$, $\mathcal{C} \supseteq \mathcal{C}'$, and $\mathcal{H} \supseteq \mathcal{H}'$, then $\mathcal{C}'$ is PAC-learnable from $\mathcal{H}'$.

☐ True ☐ False

b) Every concept class $\mathcal{C}$ is PAC-learnable from itself.

☐ True ☐ False

c) If $\mathcal{C}$ has finite VC-dimension and $\hat{c}_n^* = \arg\min_{c \in \mathcal{C}} \hat{\mathcal{R}}_n(c)$, then $\mathcal{R}(\hat{c}_n^*)$ converges in probability to $\inf_{c \in \mathcal{C}} \mathcal{R}(c)$ as $n \to \infty$.

☐ True ☐ False

2. Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{C}$ be a concept class. Assume that there is an algorithm $\mathcal{A}$ that, when given as input a sample of labeled instances of size $n$, outputs a concept $\hat{c} \in \mathcal{C}$ with the following property: For any $\epsilon > 0$, $\mathbf{P}\left(\mathcal{R}(\hat{c}) - \inf_{c \in \mathcal{C}} \mathcal{R}(c) > \epsilon\right) \leq \exp(-\epsilon n)$. Prove that $\mathcal{C}$ is PAC-learnable from itself.

**3 pts**

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .