Université Nice Sophia-Antipolis
Statistical learning, 2021-2022
`https://math.unice.fr/~rubentha/cours.html`

Final examination (MPA & MathMods), B

Documents and calculators forbidden. Give back the subject with your copy (+0.5 points!).
Duration: 2h30.

Part 1. Multiple choice questions (5 points, write the answers on the examination copy, without justification (this is a quiz). One answer per question, 0.5 point for a correct answer (zero point otherwise))

(1) What type of machine learning algorithm makes predictions when you have a set of input data and you know the possible responses?
  (a) Unsupervised learning.
  (b) Supervised learning.
  (c) Data prophecy.
  (d) Supervisory logic.
(2) When would you reduce dimensions in your data?
  (a) When your data set is larger than 500 GB.
  (b) When the data comes from sensors.
  (c) When you have a large set of features with similar characteristics.
  (d) When you are using a Linux machine.
(3) What is principal component analysis?
  (a) A linear feature transformation technique for reducing data dimensionality.
  (b) A feature selection technique that adds or removes features to optimize prediction accuracy.
  (c) A clustering technique that partitions data into mutually exclusive clusters.
  (d) A predictive technique that identifies a better set of parameters.
(4) What kind of table compares classifications predicted by the model with the actual class labels?
  (a) Chaos table.
  (b) Prediction plot.
  (c) Confusion matrix.
  (d) Residual plot.
(5) _________ is a disadvantage of decision trees?
  (a) Decision trees are robust to outliers.
  (b) Decision trees are prone to be overfit.
  (c) Both A and B.
  (d) None of the above.
(6) You work for a power company that owns hundreds of thousands of electric meters. These meters are connected to the internet and transmit energy usage data in real-time. Your supervisor asks you to direct project to use machine learning to analyze this usage data. Why are machine learning algorithms ideal in this scenario?
  (a) The algorithms will improve the wireless connectivity.
  (b) The algorithms would help the meters access the internet.
  (c) The algorithms would help your organization see patterns of the data.
(7) To predict a quantity value. use ___.
  (a) Clustering.
  (b) Classification.
  (c) Dimensionality reduction.
  (d) Regression.
(8) Why is naive Bayes called naive?

(a) It naively assumes that you will have no data.
(b) It naively assumes that the predictors are independent from one another
(c) It does not even try to create accurate predictions
(d) It naively assumes that all the predictors depend on one another..

(9) What is one reason not to use the same data for both your training set and your testing set?
(a) You will pick the wrong algorithm.
(b) You will almost certainly under-fit the model.
(c) You will almost certainly over-fit the model.
(d) You might not have enough data for both.

(10) You are working on a project that involves clustering together images of different dogs. You take image and identify it as your centroid image. What type machine learning algorithm are you using?
(a) K-means clustering.
(b) Centroid reinforcement.
(c) K-nearest neighbour.
(d) Binary classification.

Part 2. Mathematics exercises (all exercises are independent)

Exercise 1. (5 points)We are interested in estimating parameters α, c. We have independent observations $x_1, \dots, x_n$ ($n \in \mathbb{N}^*$), all of density

$$x \in \mathbb{R} \mapsto \text{Pareto}(x|\alpha, c) = \frac{\alpha c^\alpha}{x^{\alpha+1}} \mathbb{1}_{x>c} \,.$$

(1) We suppose the prior on α, c is $p(\alpha, c) = \mathbb{1}_{\alpha,c>0}$. Compute the posterior $p(\alpha, c|x_1, \dots, x_n)$.
(2) Compute $p(\alpha|c, x_1, \dots, x_n)$.
(3) Compute $p(\alpha|c, x_1, \dots, x_n)$.

Exercise 2. (5 points) We consider the space $\mathbb{R}^p$ with the euclidean distance ($p \in \mathbb{N}^*$). We have N points in $\mathbb{R}^p$, uniformly distributed in the ball of centre 0 and radius 1/2, and independent ($N \in \mathbb{N}^*$). The volume of the ball of center 0 and radius 1/2 is 1. For any r, the volume of the ball of radius r is $v_p r^P$, for some constant v_p. Let R be the distance from the origin to its nearest neighbour amongst the N points. Show that

$$\text{median}(R) = v_p^{-1/p} \left(1 - \left(\frac{1}{2} \right)^{1/N} \right)^{1/p} \,.$$

Remember that median(R) is the number m such that $\mathbb{P}(R > m) = \frac{1}{2}$.

Exercise 3. (5 points) We have vectors $x^{(1)}, \dots, x^{(N)}$ in $\mathbb{R}^D$ ($N > D$). We have $t_1, \dots, t_N$ in $\mathbb{R}$. We are interested in

$$\widehat{w} = \underset{w \in \mathbb{R}^D}{\arg\min} \sum_{i=1}^{N} (t_i - w^T x^{(i)})^2 \,.$$

We set

$$x^{(i)} = \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_D^{(i)} \end{pmatrix} , \forall i \,,$$

$$X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_D^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(N)} & x_2^{(N)} & \dots & x_D^{(N)} \end{bmatrix} \,.$$

(1) Show that (for all w)

$$\sum_{i=1}^{N}(t_i - w^T x^{(i)})^2 = \left(\begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{pmatrix} - Xw\right)^T \left(\begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{pmatrix} - Xw\right).$$

(2) We set

$$\mathbf{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{pmatrix}, \ \mathcal{L}(w) = \sum_{i=1}^{N}(t_i - w^T x^{(i)})^2 .$$

Show that the gradient of $\mathcal{L}$ is

$$\nabla\mathcal{L}(w) = 2(X^T X)w - 2X^T\mathbf{t} .$$

(3) Prove that $X^T X$ is invertible.

(4) Find the absolute minimum of $\mathcal{L}$.