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

Final examination (IM), B

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

Part 1. Multiple choice questions (10 points, write the answers on the examination copy, without justification (this is a quiz). One answer per question, one 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) Supervised learning.
  (b) Supervisory logic.
  (c) Deep learning.
  (d) Unsupervised learning.

(2) When would you reduce dimensions in your data?
  (a) When the data comes from sensors.
  (b) When your data set is larger than 500 GB.
  (c) When you are using a Linux machine.
  (d) When you have a large set of features with similar characteristics.

(3) What does a classification model do?
  (a) Predicts real number responses such as changes in temperature, date, or time.
  (b) Clusters responses in groups based on similarity, to find patterns.
  (c) Assigns data to a predefined category.
  (d) Compares predicted data classifications to the actual class labels in the data .

(4) What is principal component analysis?
  (a) A feature selection technique that adds or removes features to optimize prediction accuracy.
  (b) A clustering technique that partitions data into mutually exclusive clusters.
  (c) A linear feature transformation technique for reducing data dimensionality.
  (d) A predictive technique that identifies a better set of parameters.

(5) What is overfitting?
  (a) When a predictive model is accurate but takes too long to run.
  (b) When you apply a powerful deep learning algorithm to a simple machine learning problem.
  (c) When the model learns specifics of the training data that can't be generalized to a larger data set.
  (d) When you perform hyper-parameter tuning and performance degrades.

(6) ___________ is the machine learning algorithms that can be used with unlabeled data.
  (a) Regression algorithms.
  (b) Clustering algorithms.
  (c) All of the above.

(7) _________ 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.

(8) ____ looks at the relationship between predictors and your outcome.
  (a) Regression analysis.

(b) K-means clustering.
(c) Big data.
(d) Unsupervised learning.

(9) 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 would help the meters access the internet.
(b) The algorithms will improve the wireless connectivity.
(c) The algorithms would help your organization see patterns of the data.

(10) You work for an insurance company. Which machine learning project would add the most value for the company?
(a) Create an artificial neural network that would host the company directory.
(b) Use machine learning to better predict risk.
(c) Create an algorithm that consolidates all of your Excel spreadsheets into one data lake.
(d) Use machine learning and big data to research salary requirements.

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 $p(c|\alpha, x_1, \dots, x_n)$.

Exercise 2. (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) We suppose that $X^T X$ is invertible. Find the absolute minimum of $\mathcal{L}$.