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

Final examination (MPA & MathMods), A

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) Supervisory logic.
- (b) Supervised learning.
- (c) Unsupervised learning.
- (d) Deep learning.

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

(3) 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.

(4) 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 you perform hyper-parameter tuning and performance degrades.
- (d) When the model learns specifics of the training data that can't be generalized to a larger data set.

(5) What kind of table compares classifications predicted by the model with the actual class labels?
- (a) Chaos table.
- (b) Prediction plot.
- (c) Residual plot.
- (d) Confusion matrix.

(6) Application of Machine learning is _ _ _ _ _ _.
- (a) Sentimental analysis.
- (b) E-mail filtering.
- (c) Face recognition.
- (d) All of the above.

(7) _ _ _ _ _ _ _ _ _ is a disadvantage of decision trees?
- (a) Decision trees are prone to be overfit.
- (b) Decision trees are robust to outliers.
- (c) Both A and B.
- (d) None of the above.

(8) _ _ _ _ looks at the relationship between predictors and your outcome.

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

(9) What is an example of a commercial application for a machine learning system?
(a) A data entry system.
(b) A data warehouse system.
(c) A massive data repository.
(d) A product recommendation system.

(10) 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.

Part 2. Mathematics exercises (all exercises are independent)

Exercise 1. (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 2. (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 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 $\mathcal{L}$.