Université Nice Sophia-Antipolis
Statistical leaarning, 2020-2021
`https://math.unice.fr/~rubentha/cours.html`

Final examination, (MPA $\cup$ MATHMODS), A

Documents and calculators forbidden. Give back the subject with your copy.

1. Quiz (*Write the answers on the examination copy, without justification (this is a quiz). One answer per question. One point per good answer (zero point for a bad answer).*

(1) Assuming that you have a very large training set, which of the following algorithms do you think can be parallelized using map-reduce and splitting the training set across different machines?
(a) A neural network trained using batch gradient descent.
(b) Linear regression trained using batch gradient descent.
(c) Logistic regression trained using stochastic gradient descent.

(2) I plot the loss error during the learning phase of my neural network (the spiky line in Figure 1.1). The loss error is the sum of the squares of the distances between what the network produces and the output (which is known on the training set). I also plot the loss error on a validation set (a set for which I know the output) (the smooth line in Figure 1.1).

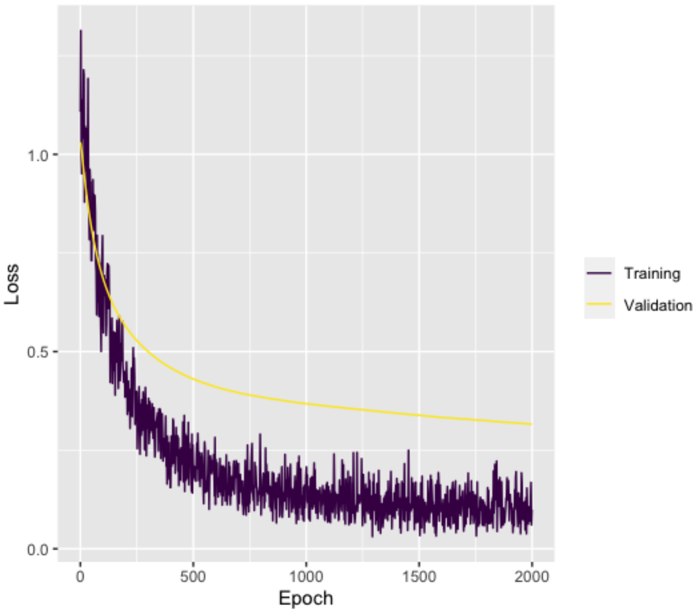

Figure 1.1. Loss error

(a) I am over-fitting.
(b) I am under-fitting.
(c) None of the above, I am good.

(3) It is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. This clustering only works with numeric data. Which clustering has been discussed here?
(a) K-nearest neighbour

(b) K-means
(c) K-clustering

(4) We do a linear regression. Which of the following is true in the context of regularization/penalization?
(a) Ridge regression is also called $L1$ regularization.
(b) Ridge regression can reduce the parameters to zero.
(c) Lasso regression can reduce the parameters to a large extent but not to zero.
(d) None of the above.

(5) Imagine, you are solving a classification problems with highly imbalanced class. The majority class is observed 99% of times in the training data. Your model has 99% accuracy after taking the predictions on test data. Which of the following is true in such a case?
(a) I should find another performance measure.
(b) My model is bad.
(c) Both (a) and (b).

(6) To find the minimum or the maximum of a function, we set the gradient to zero because
(a) The value of the gradient at extrema of a function is always zero.
(b) Depends on the type of problem.
(c) None of the above.

(7) Which of the following statements about regularization is not correct? (We add a penalization/regularization term multiplied by λ.)
(a) Using too large a value of lambda can cause your model to under-fit the data.
(b) Using too large a value of lambda can cause your model to over-fit the data.
(c) Using a very large value of lambda cannot hurt the performance of your model.

(8) Which of the following is NOT supervised learning?
(a) Principal Component Analysis.
(b) Linear Regression.
(c) Naïve Bayes Classifier.

(9) What is the purpose of performing cross-validation?
(a) To assess the predictive performance of the models .
(b) To judge how the trained model performs outside the sample on test data.
(c) Both 1 and 2.

(10) Which of the following is an example of a deterministic algorithm?
(a) Principal Component Analysis.
(b) K-Means.
(c) None of the above.

2. More difficult questions (five points for each exercise)

Exercise 1. We have a random variable Θ in $\mathbb{R}^+$ with density: $x \in \mathbb{R}^+ \mapsto e^{-x}$. We have a Markov chain $(X_k)_{k\geq 0}$ such that

- $X_0 = 0$
- $$X_{k+1} = \begin{cases} X_k & \text{with probability 1-}\Theta\,, \\ X_k + U_k & \text{with probability } \Theta/2\,, \\ X_k - U_k & \text{with probability } \Theta/2\,, \end{cases}$$

  where the (U_k) are independent, with the same law $\mathcal{N}(0,1)$.

We have observations (for all k)

$$Y_k = X_k + V_k$$

where the (V_k) are independent (and independent of the (U_k)), with the same law $\mathcal{N}(0,1)$.

For random variables A, B, we use the following notation: $\mathbb{P}(A=a)$ is the density of A taken in a, $\mathbb{P}(A=a|B=b)$ is the density of $\mathcal{L}(A|B=b)$ taken in b.

We suppose we have observations $Y_0 = y_0, Y_1 = y_1, \dots, Y_n = y_n$ (for some n) that are fixed.

(1) Using the Bayes formula, show that

$$\mathbb{P}((X_k)_{0\leq k\leq n}=(x_k)_{0\leq k\leq n},\Theta=\theta|(Y_k)_{0\leq k\leq n}=(y_k)_{0\leq k\leq n})$$

is proportional to

$$\mathbb{P}((Y_k)_{0\leq k\leq n}=(y_k)_{0\leq k\leq n}|(X_k)_{0\leq k\leq n}=(x_k)_{0\leq k\leq n},\Theta=\theta)$$
$$\times\,\mathbb{P}((X_k)_{0\leq k\leq n}=(x_k)_{0\leq k\leq n}|\Theta=\theta)\mathbb{P}(\Theta=\theta)\,.$$

(2) Propose a MCMC scheme in order to get a sample of the law $\mathbb{P}(\Theta|(Y_k)_{0\leq k\leq n}=(y_k)_{0\leq k\leq n})$. You do not have to prove the convergence but you must specify the following.
  (a) State space of your MCMC.
  (b) Proposal kernel.
  (c) Target law.

Exercise 2. We want to minimize

$$H\,:\,z\in\mathbb{R}^d\mapsto-\sum_{k=1}^{d}y_k\log(z_k)$$

where $d\in\mathbb{N}^*$ and $y_1,\dots,y_d$ are fixed (in $\mathbb{R}^+$ and such that $y_1+\dots+y_d=1$). We want to do this minimization under the constrain

$$z_1\geq 0,\dots,z_d\geq 0\ \text{ and }\ z_1+\dots+z_d=1\,.$$

We set $F(z)=z_1+\dots+z_d$.

(1) Compute the gradient of H.
(2) Compute the gradient of F.
(3) Find a candidate for the minimum.
(4) Show that this candidate is an absolute minimum.