Exercises - Machine Learning
Master 1 Informatique Fondamentale ENS Lyon

Yohann De Castro & Aurélien Garivier

draft version of January 2020

Hands-on Session 1: Statistics 101

Friday 17th January, 2020

*** Exercise 1 Maximum Likelihood Estimators

For a given sample size $n \geq 1$, compute the Maximum Likelihood Estimators in the models $\big(\mathcal{X}^n, \{Q_\theta^{\otimes n}\}\big)$ in each of the following cases.

1. $\mathcal{X} = \mathbb{R}$, $Q_\theta = \mathcal{N}(\theta, \sigma^2)$, where σ is a known parameter.
2. $\mathcal{X} = \mathbb{R}$, $Q_\theta = \mathcal{N}(\mu, \sigma^2)$, where $\theta = (\mu, \sigma^2) \in \mathbb{R} \times [0, +\infty)$.
3. $\mathcal{X} = \{0, 1\}$, $Q_\theta = \mathcal{B}(\theta)$.
4. $\mathcal{X} = \mathbb{R}^+$, $Q_\theta = \mathcal{U}\big([0, \theta]\big)$.
5. $\mathcal{X} = \mathbb{R}$, $Q_\theta = \mathcal{E}(\theta)$.
6. $\mathcal{X} = \mathbb{R}$, $Q_\theta = \mathcal{L}(\theta)$ the *Laplace* distribution centered at θ, which has density $f_\theta(x) = \exp(-|x - \theta|)/2$.

   Whenever possible, compute the quadratic risks of the obtained estimators.

*** Exercise 2 Confidence Intervals

In all the following models, with sample size $n \geq 1$, propose a confidence interval for θ. Precise whether it is asymptotic or not.

1. $\mathcal{X} = \mathbb{R}$, $Q_\theta = \mathcal{N}(\theta, \sigma^2)$, where σ is a known parameter.
2. $\mathcal{X} = \{0, 1\}$, $Q_\theta = \mathcal{B}(\theta)$.

* 3. $\mathcal{X} = \mathbb{R}^+$, $Q_\theta = \mathcal{U}\big([0, \theta]\big)$.

** 4. $\mathcal{X} = \mathbb{R}$, $Q_\theta = \mathcal{L}(\theta)$.

** Hands on 1 Mean or Median?

We consider an odd sample size $n = 2k - 1$, and the two following models:

$$\mathcal{M}_1 = \Big(\mathbb{R}^n, \big\{\mathcal{N}(\mu, 1)^{\otimes n} : \mu \in \mathbb{R}\big\}\Big)\,,$$
$$\mathcal{M}_2 = \Big(\mathbb{R}^n, \big\{\mathcal{L}(\mu)^{\otimes n} : \mu \in \mathbb{R}\big\}\Big)\,.$$

For each model, give the properties of the two following estimators:

$$\hat{\mu}_n = \frac{X_1 + \cdots + X_n}{n} \text{ the sample mean, and}$$
$$\tilde{\mu}_n = X_{(k)} \text{ the sample median.}$$

Numerically estimate the quadratic risk of each estimator in each model.
Comment the results.

*** Hands on 2 Linear Regression with scikitlearn

Experiment linear regression with scikitlearn on the reference example `https://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html`.

You will need to load a dataset made of n = 442 diabetes patients, with for each patient the disease progression one year after baseline, and 10 variables: age, sex, body mass index, average blood pressure, and six blood serum measurements.

Try to answer the following question: what is the best linear model for predicting the response given the features, and how reliable are the predictions?

Hands-on Session 2: Clustering

Friday 24th January, 2020

**** Exercise 3 On the Consistency of K-Means

Let us consider n points $X_1, \dots, X_n$ in $\mathbb{R}^p$. The K-means algorithm seeks to minimize over all partitions $G = (G_1, \dots, G_K)$ of $\{1, \dots, p\}$ the criterion

$$\mathrm{crit}(G) = \sum_{k=1}^{K} \sum_{a \in G_k} \|X_a - \bar{X}_{G_k}\|^2 \quad \text{with} \quad \bar{X}_{G_k} = \frac{1}{|G_k|} \sum_{b \in G_k} X_b.$$

1. *(Symmetrization)* To analyse the K-means, it is useful to symmetrize the criterion. Prove the two equalities

$$\begin{aligned} \mathrm{crit}(G) &= \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{a,b \in G_k} \langle X_a, X_a - X_b \rangle \\ &= \frac{1}{2} \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{a,b \in G_k} \|X_a - X_b\|^2. \end{aligned}$$

2. *(Independent observations)* We assume now that the observations are random and independent. We write $\mu_a \in \mathbb{R}^p$ for the expectation of X_a so that $X_a = \mu_a + \varepsilon_a$ with $\varepsilon_1, \dots, \varepsilon_n$ centered and independent. We define $v_a = \mathrm{trace}(cov(X_a))$.
Check that the expected value of the criterion is

$$\mathbb{E}[\mathrm{crit}(G)] = \frac{1}{2} \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{a,b \in G_k} \left(\|\mu_a - \mu_b\|^2 + v_a + v_b\right) \mathbf{1}_{a \neq b}.$$

What is the value of $\mathbb{E}[\mathrm{crit}(G)]$ when all the within-group variables have the same mean?

3. *(Mixture model)* We assume now that there exists a partition $G^* = (G_1^*, \dots, G_K^*)$ such that within-group variables have the same mean and the same volume. More precisely, we assume that there exists $m_1, \dots, m_K \in \mathbb{R}^p$ and $\gamma_1, \dots, \gamma_K > 0$ such that $\mu_a = m_k$ and $v_a = \gamma_k$ for all $a \in G_k^*$ and $k = 1, \dots, K$.
Below, we investigate under which condition the expected value of the Kmeans criterion is minimum in G^*.
   a) What is the value of $\mathbb{E}[\mathrm{crit}(G^*)]$?
   b) In the special case where $\gamma_1 = \dots = \gamma_K = \gamma$, which partition $G = (G_1, \dots, G_K)$ minimizes $\mathbb{E}[\mathrm{crit}(G)]$?
   c) We assume now that we have $K = 3$ groups of size s (with s even),

$$m_1 = (1, 0, 0)^T, \quad m_2 = (0, 1, 0)^T, \quad m_3 = (0, 1 - \tau, \sqrt{1 - (1 - \tau)^2})^T,$$

   with $\tau > 0$, and

$$\gamma_1 = \gamma_+, \quad \gamma_2 = \gamma_3 = \gamma_-.$$

   What is the value of $\|m_2 - m_3\|^2$?
   d) Compute $\mathbb{E}[\mathrm{crit}(G^*)]$.
   e) Let us define G' obtained by splitting G_1^* into two groups G_1', G_2' of equal size $s/2$ and by merging G_2^* and G_3^* into a single group G_3' of size $2s$. Check that

$$\mathbb{E}[\mathrm{crit}(G')] = s(\gamma_+ + 2\gamma_- + \tau) - (2\gamma_+ + \gamma_-).$$

f) When do we have $\mathbb{E}[\text{crit}(G^*)] < \mathbb{E}[\text{crit}(G')]$?
g) What is the take home message?

Conversely, in the general mixture model, we can check that if

$$\min_{j \neq k} \|m_j - m_k\|^2 > 2\,\frac{\max_k \gamma_k - \min_k \gamma_k}{\min_k |G_k^*|}$$

then $\mathbb{E}[\text{crit}(G^*)] < \mathbb{E}[\text{crit}(G)]$ for all partitions $G = (G_1, \ldots, G_K)$ not equal to G^*.

*** Hands on 3 Clustering of text

See attached notebook.

Hands-on Session 3: Dimensionality Reduction

Friday 31th January, 2020

**** Exercise 4 Computing the largest eigenvalue

Let $A \in \mathcal{M}_\backslash(\mathbb{R})$ be a symmetric, positive matrix, and let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ be its eigenvalues. For each $i \in \{2, \ldots, n\}$ let $v^i \in \mathbb{R}^n$ be such that $\|v^i\| = 1$ be such that $Av^i = \lambda_i v^i$:

We assume that $\lambda_1 > \lambda_2$. The goal of this exercise is to analyze a probabilistic algorithm approximating $v := v^1$. The algorithm, called *power iteration*, relies on the following induction: $u_0 = \left[\frac{\epsilon_1}{\sqrt{n}}, \ldots, \frac{\epsilon_n}{\sqrt{n}}\right]$ where $\epsilon_i \overset{iid}{\sim} \mathcal{U}\big(\{-1, 1\}\big)$ and for all $t \geq 1$, $u_{t+1} = \frac{Au_t}{\left\|Au_t\right\|}$.

1. Show that for all $t \geq 0$, $\|u_t\| = 1$ and

$$u_t = \frac{A^t u_0}{\|A^t u_0\|} = \frac{\sum_{i=1}^n \lambda_i^t \langle u_0, v_i \rangle v_i}{\sqrt{\sum_{i=1}^n \left(\lambda_i^t \langle u_0, v_i \rangle\right)^2}} .$$

2. What are the expectation and variance of $\langle u_0, v \rangle$?
3. Denoting $Z = \langle u_0, v \rangle^2$, show that $\mathbb{E}[Z] = 1/n$ and that $\mathbb{E}[Z^2] \leq 3/n^2$.
4. Let $\delta \in (0, 1)$. Using the Cauchy-Schwartz inequality with the variables X and $\mathbb{1}\big\{X > \delta\mathbb{E}[X]\big\}$, show that for every non-negative random variable X with finite variance

$$\mathbb{P}\big(X \geq \delta\mathbb{E}[X]\big) \geq (1-\delta)^2 \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]} .$$

5. Prove that

$$\mathbb{P}\left(Z \geq \frac{1}{4n}\right) \geq \frac{3}{16} .$$

6. Show that whenever $\langle u_0, v \rangle^2 > 1/(4n)$,

$$\big|\langle u_t, v \rangle\big| = \frac{1}{\sqrt{1 + \frac{1}{\langle u_0, v \rangle^2} \sum_{i=2}^n \langle u_0, v^i \rangle^2 \left(\frac{\lambda_i}{\lambda_1}\right)^{2t}}} \geq 1 - 2n\left(\frac{\lambda_2}{\lambda_1}\right)^{2t} .$$

7. Summarize the conclusion of the two previous questions.
8. For a fixed $\epsilon > 0$, how many iterations does it take to obtain with probability at least 95% a vector u such that $\big|\langle u_t, v \rangle\big| \geq 1 - \epsilon$?

Remark: one can similarly show that with non-vanishing probability

$$\langle u_t, Au_t \rangle \geq \lambda_1 \times \frac{1-\epsilon}{1 + 4n(1-\epsilon)^{2t}} .$$

See `http://theory.stanford.edu/~trevisan/expander-online/lecture03.pdf`.

*** Hands on 4 Dimensionality Reduction for the MNIST classification problem

See attached notebook.

Hands-on Session 4: Introduction to Supervised Learning

Friday 7th February, 2020

* Exercise 5 Classification 1

Consider the binary classification problem with the following (not usual) risk

$$\ell(\widehat{y}, y) := \begin{cases} c & \text{if } \widehat{y} = 1,\ y = 0 \\ 1 & \text{if } \widehat{y} = 0,\ y = 1 \\ 0 & \text{otherwise} \end{cases}$$

1. Compute the classification risk of a rule g, namely

$$L(g) := \mathbb{E}[\ell(g(X), Y)]$$

2. Show that the optimal Bayes rule $f^\star$ is given by

$$f^\star(x) = \mathbb{1}_{\eta(x) \geq \frac{c}{1+c}}\,,$$

   where $\eta(x) := \mathbb{E}[Y|X = x] = \mathbb{P}(Y = 1|X = x)$.

** Exercise 6 Classification 2

Consider the binary classification problem. Let g and g' be two classification rules. Let L be the standard 0/1 loss ($c = 1$ in the aforementioned exercise).

1. Show that

$$|L(g) - L(g')| \leq \mathbb{P}(g(X) \neq g'(X))$$

2. Show that

$$L(g) = \mathbb{E}[\mathbb{1}_{\{g(X)\neq 1\}}(2\eta(X) - 1) + (1 - \eta(X))]\,,$$

   where $\eta(x) := \mathbb{E}[Y|X = x] = \mathbb{P}(Y = 1|X = x)$.

3. Show that

$$|L(g) - L(g')| \leq \mathbb{E}[|2\eta(X) - 1|\mathbb{1}_{\{g(X)\neq g'(X)\}}]\,.$$

Now, for two sets A and B, we denote $A\Delta B := (A \cap B^c) \cup (A^c \cap B)$ their symmetric difference.

4. Show that

$$L(g) - L^\star = \mathbb{E}[|2\eta(X) - 1|\mathbb{1}_{G\Delta G^\star}(X)]$$

   where $L^\star$ is the optimal risk (the infimum), $G = g^{-1}(\{1\})$ and $G^\star = (g^\star)^{-1}(\{1\})$ with $g^\star$ the optimal Bayes classifier. In particular, note that $G^\star = \{x \in \mathcal{X} \ :\ \eta(x) \geq 1/2\}$.

In practice, we may have access to an estimation $\widehat{\pi}_0, \widehat{\pi}_1, \widehat{p}_0, \widehat{p}_1$ of

$$\begin{aligned}
\pi_0 &= \mathbb{P}(Y = 0), \\
\pi_1 &= \mathbb{P}(Y = 1), \\
p_0(x) &= \mathbb{P}(X = x|Y = 0), \\
p_1(x) &= \mathbb{P}(X = x|Y = 1),
\end{aligned}$$

and we may denote

$$\widehat{\eta}(x) = \frac{\widehat{\pi}_1 \widehat{p}_1(x)}{\widehat{\pi}_0 \widehat{p}_0(x) + \widehat{\pi}_1 \widehat{p}_1(x)}\,,$$

the deduced estimation of $\eta(x)$. Consider the following rule of classification

$$\widehat{g}(x) = \mathbb{1}_{\{\widehat{\eta}(x) \geq 1/2\}}\,.$$

5. Show that

$$L(\widehat{g}) - L^\star \leq \int_{\mathcal{X}} \sum_{k=0}^{1} |\pi_k p_k(x) - \widehat{\pi}_k \widehat{p}_k(x)| \mathrm{d}\mu(x)\,,$$

   where μ is the law of X.

*** Exercise 7 An analysis of the Nearest-Neighbour Algorithm

We consider the problem of binary classification ($\mathcal{Y} = \{0, 1\}$) on the feature set $\mathcal{X} = [0, 1[^d$ with the nearest-neighbour method: if the training set is $S_n = \big\{(X_1, Y_1), \ldots, (X_n, Y_n)\big\}$, then for all $x \in \mathcal{X}$ we define

$$I(x) = \operatorname*{arg\,min}_{1 \leq i \leq n} \|x - X_i\| \quad \text{and} \quad \hat{h}_n(x) = Y_{I(x)}\,.$$

The objective of this exercise is to prove a bound on the risk $R\big(\hat{h}_n\big) = \mathbb{E}_{S_n}\Big[\mathbb{P}_{X,Y}\big(\hat{h}_n(X) \neq Y\big)\Big]$ of $\hat{h}_n$, under the assumption that $\eta : x \mapsto \mathbb{P}(Y = 1|X = x)$ is c-Lipschitz continuous for a positive constant c:

$$\forall x, x' \in \mathcal{X}, \big|\eta(x) - \eta(x')\big| \leq c\big\|x - x'\big\|\,.$$

.

1. Show that $h^* : x \mapsto \mathbb{1}\{\eta(x) \geq 1/2\}$ is a Bayes classifier and has loss $L^* = \mathbb{P}\big(h^*(X) \neq Y\big) = \mathbb{E}\Big[\min\big(\eta(X), 1 - \eta(X)\big)\Big]$.
2. Show that if $Z_1 \sim \mathcal{B}(p)$ and $Z_2 \sim \mathcal{B}(q)$ are two independent variables, then $\mathbb{P}(Z_1 \neq Z_2) \leq 2\min(p, 1-p) + |p - q|$.
3. Show that

$$R_n\big(\hat{h}_n\big) = \mathbb{E}\Big[\mathbb{E}\big[\mathbb{1}\{Y \neq Y_{I(X)}\}\big|X, X_1, \ldots, X_n\big]\Big]\,.$$

4. Prove that

$$\mathbb{E}\Big[\mathbb{1}\{Y \neq Y_{I(X)}\}|X, X_1, \ldots, X_n\Big] \leq 2\min\big(\eta(X), 1 - \eta(X)\big) + c\,\big\|X - X_{I(X)}\big\|\,.$$

5. We consider the partition $\mathcal{C}$ of $\mathcal{X}$ into $|\mathcal{C}| = T^d$ cells of diameter $\sqrt{d}/T$:

$$\mathcal{C} = \left\{\left[\frac{j_1 - 1}{T}, \frac{j_1}{T}\right[\times \cdots \times \left[\frac{j_d - 1}{T}, \frac{j_d}{T}\right[, \quad 1 \leq j_1, \ldots, j_d \leq T\right\}\,.$$

   Show that

$$\big\|X - X_{I(X)}\big\| \leq \sum_{c \in \mathcal{C}} \mathbb{1}\{X \in c\} \left(\frac{\sqrt{d}}{T} \mathbb{1} \bigcup_{i=1}^{n} \{X_i \in c\} + \sqrt{d}\mathbb{1} \bigcap_{i=1}^{n} \{X_i \notin c\}\right)\,.$$

6. For every cell $c \in \mathcal{C}$ of probability $p_c = \mathbb{P}(X \in c)$, prove that

$$\mathbb{P}\left(\big\{X \in c\big\} \cap \bigcap_{i=1}^{n} \big\{X_i \notin c\big\}\right) \leq p_c\, e^{-n\, p_c} \leq \frac{1}{e\,n}\,.$$

7. Prove that
$$\mathbb{E}\big[\|X - X_{I(X)}\|\big] \le \frac{\sqrt{d}}{T} + \frac{\sqrt{d}T^d}{e\,n}\ .$$
8. Conclude:
$$R_n\big(\hat{h}_n\big) \le 2L^* + \frac{3c\sqrt{d}}{n^{1/(d+1)}}\ .$$

Hands-on Session 5: Cross-Validation and Model Selection

Friday 14th February, 2020

**** Exercise 8 Model Selection

This exercice is important as it presents, in a simple framework, the notion of regularization.

Experience: Consider $\mathbf{X} \sim \mathcal{N}_p(\mu^0, \Sigma)$ a Gaussian vector of size p, mean $\mu^0 \in \mathbb{R}^p$, and variance Σ a positive semidefinite (psd) matrix. For sake of simplicity we assume that $\Sigma = \sigma^2 \mathrm{Id}_p$ where $\sigma > 0$ is known. We observe $X_1, \ldots, X_n \sim \mathbf{X}$ i.i.d. vectors.

Task: Let V be a known orthonormal matrix (i.e. $VV^\top = V^\top V = \mathrm{Id}_p$) and denote

$$\underbrace{\mathrm{Span(V_1)}}_{E_1} \subset \cdots \subset \underbrace{\mathrm{Span(V_k)}}_{E_k} \subset \cdots \subset \underbrace{\mathrm{Span(V)}}_{\mathbb{R}^p}$$

where V_k is the $p \times k$ matrix obtained from V keeping the k first columns. In particular

$$\Pi_k := V_k V_k^\top \text{ is the orthogonal projection on } E_k$$

Assume that $\mu^0 = V\theta^0$ for some unknowns $k^0 \in [p]$ and $\theta^0 = (\theta^0_1, \ldots, \theta^0_{k^0}, 0, \ldots, 0) \in \mathbb{R}^{k^0} \times \{0\}^{p-k^0}$ with $\theta^0_{k^0} \neq 0$. Note that $\theta^0_k = 0$ for $k > k^0$. The goal is to recover a good approximation $\widehat{\mu}$ of μ^0, where $\widehat{\mu}$ can be any measurable function of $(X_1, \ldots, X_n)$. This basic framework depicts important cases where one seeks to recover the decomposition of the "classifier" in some known orthonormal basis V.

Performance: Performance is measured by the following risk

$$\mathcal{R}(\widehat{\mu}) := \mathbb{E}\|\mathbf{X} - \widehat{\mu}\|_2^2 - \sigma^2 p\,,$$

where the expectation is taken with respect to $\mathbf{X}, X_1, \ldots, X_n$ which are i.i.d. vectors and such that $\widehat{\mu} = \widehat{\mu}(X_1, \ldots, X_n)$.

1. Show that for all measurable function $\widehat{\mu}(X_1, \ldots, X_n)$ it holds

$$\mathcal{R}(\widehat{\mu}) = \mathbb{E}\|\mu^0 - \widehat{\mu}\|_2^2\,,$$

   where the expectation is taken with respect to $X_1, \ldots, X_n$.

Strategies: We start with some very elementary questions.

2. Compute the law of $V^\top \mathbf{X}/\sigma$.
3. Prove that the problem can be equivalently reduced to the case $V = \mathrm{Id}_p$ and $\sigma = 1$. We will assume it from now.

A first strategy, that matches what you may have seen in Statistics before, goes by using the "Empirical Risk Minimizer" (ERM). Indeed, the risk function $\mathcal{R}(\widehat{\mu})$ is not observed since it depends on the target μ^0 but an empirical version of the risk may be computed as

$$\mu \mapsto \mathcal{R}_n(\mu) := \frac{1}{n}\sum_{k=1}^{n} \|X_k - \mu\|_2^2 - \sigma^2 p\,.$$

4. Compute the minimum $\widehat{\mu}^{\mathrm{ERM}}$ of the empirical risk $\mathcal{R}_n$.
5. Compute its risk $\mathcal{R}(\widehat{\mu}^{\mathrm{ERM}})$.

Now, assume that someone (referred to as the "oracle") reveals you the true value of k^0.

6. Can you build $\widehat{\mu}^{\text{oracle}}$ which is the Best Linear Unbiased Estimator (BLUE) of the mean μ^0?

7. When $k^0 < p$, show that

$$\mathcal{R}(\widehat{\mu}^{\text{oracle}}) = \frac{k^0}{n} < \frac{p}{n} = \mathcal{R}(\widehat{\mu}^{\text{ERM}})\,.$$

Of course, we don't know k^0. The strategy is then to "penalize" the Empirical Risk so as to reduce its "bias".

8. Compute the variance of $\widehat{\mu}^k := \Pi_k \bar{X}$, where $\bar{X}$ is the empirical mean

$$\bar{X} := \frac{1}{n}\sum_{k=1}^{n} X_k\,.$$

9. Show that

$$\mathcal{R}_n(\widehat{\mu}^k) = \|\mu^0 - \widehat{\mu}^k\|_2^2 + 2\langle \bar{X} - \mu^0, \mu^0 - \Pi_k \bar{X}\rangle + \frac{1}{n}\sum_{j=1}^{n} \|X_j - \mu^0\|_2^2 - \sigma^2 p$$

10. Consider the following penalized estimator

$$\widehat{k} := \arg\min_{k\in[p]} \Big\{\mathcal{R}_n(\widehat{\mu}^k) + \lambda\frac{k}{n}\Big\}\,,$$

where $\lambda > 0$ is a tuning parameter. Our penalized estimator is then $\widehat{\mu}^{\text{pen}} := \widehat{\mu}^{\widehat{k}}$. We won't study into details this estimator, this is the core of the course "Model Selection". We rather investigate some heuristics here and elementary manipulations. Prove that for all $0 < \alpha < 1$, it holds

$$\|\mu^0 - \widehat{\mu}^{\widehat{k}}\|_2^2 \leq \frac{1}{1-\alpha} \inf_k \Big\{\|\mu^0 - \widehat{\mu}^k\|_2^2 + \lambda\frac{k}{n}\Big\} + \alpha^{-1}\mathcal{O}_{\mathbb{P}}(\frac{k^0}{n}) + Z$$

where $Z = \sup_l \big(\alpha^{-1}\|\Pi_l \bar{X} - \mu^0\|_2^2 - \lambda\frac{l}{n}\big)$. This last random variable can be shown to be $\mathcal{O}_{\mathbb{P}}(1/n)$. It gives the idea that

$$\|\mu^0 - \widehat{\mu}^{\widehat{k}}\|_2^2 \leq (1 + o(1)) \inf_k \Big\{\|\mu^0 - \widehat{\mu}^k\|_2^2 + \lambda\frac{k}{n}\Big\} + \mathcal{O}_{\mathbb{P}}(1/n)\,.$$

which is called a "sharp oracle inequality".

Hint: $\langle u, v\rangle \leq \alpha\|u\|_2^2 + \alpha^{-1}\|v\|_2^2$ for all $\alpha > 0$.

Hands-on Session 6: Empirical Risk Minimization, Linear Separators

Friday 21th February, 2020

**** Exercise 9 Perceptron with margin

In this exercise, we consider binary classification in $\mathcal{X} = \mathbb{R}^d$ with label set $\mathcal{Y} = \{\pm 1\}$: the sample is $\big((x_1, y_1), \dots, (x_n, y_n)\big) \in \big(\mathbb{R}^d \times \{\pm 1\}\big)^n$. We assume that the data is linearly separable, and even that a positive *margin*

$$\gamma = \sup_{w \in \mathbb{R}^d : \|w\| = 1} \min_{1 \le i \le n} \frac{y_i \langle w, x_i \rangle}{\|x_i\|}$$

is known and can be used in the algorithm. The aim of the *Perceptron with margin* algorithm is to find a linear separator with almost optimal margin. The aim of the questions 1-6 is to prove that the Perceptron-with-margin algorithm below achieves margin at least $\gamma/2$ in at most $12/\gamma^2$ iterations.

Algorithm: Perceptron-with-margin γ

Input: margin γ
Data: training set $(x_1, y_1), \dots, (x_n, y_n)$
1 $w_0 \leftarrow (0, \dots, 0)$
2 $t \ge 0$
3 **while** $\exists i_t : y_{i_t} \langle w_t, x_{i_t} \rangle \le \frac{\gamma}{2} \|x_{i_t}\| \|w_t\|$ **do**
4 $\quad w_{t+1} = w_t + y_{i_t} \frac{x_{i_t}}{\|x_{i_t}\|}$
5 $\quad t \leftarrow t + 1$
6 **return** w_t

1. Justify the existence of $w^* \in \mathbb{R}^d$ such that $\|w^*\| = 1$ and
$$\forall 1 \le i \le n, \quad \frac{y_i \langle w^*, x_i \rangle}{\|x_i\|} \ge \gamma\ .$$
2. In this question and the following, t is a positive integer for which the condition to continue the while loop of the algorithm (line 3) is satisfied. Prove that $\langle w^*, w_t \rangle \ge \gamma t$.
3. Prove that
$$\|w_{t+1}\|^2 \le \|w_t\|^2 + \gamma \|w_t\| + 1\ .$$
4. Show that if $\|w_t\| \ge 2/\gamma$, then
$$\|w_{t+1}\|^2 \le \left(\|w_t\| + \frac{3\gamma}{4}\right)^2\ .$$
5. Deduce that
$$\|w_t\| \le 1 + \frac{2}{\gamma} + \frac{3\gamma t}{4}\ .$$
6. Conclude.
7. For any $\eta \in (0, 1)$, give an algorithm that yields a linear separator with margin at least $(1 - \eta)\gamma$ in at most $K(\eta)/\gamma^2$ iterations, where $K(\eta)$ is a function to be specified.

*** Hands on 5 Experimenting the Perceptron Algorithm

Code a perceptron for binary classification as in the previous exercise, and show the evolution of the linear separator during the iterations.

Hands-on Session 7: AdaBoost, Ensemble Methods

Friday 28th February, 2020

*** Exercise 10 AdaBoost on binary classification

Let $(x_i, y_i)_{1\le i\le n} \in (\mathcal{X} \times \{-1, 1\})^n$ be n observations and $\mathcal{H} = \{h_1, \dots, h_M\}$ be a set of M classifiers, i.e. for all $1 \le i \le M$, : $h_i : \mathcal{X} \to \{-1, 1\}$. It is assumed that for each $h \in \mathcal{H}$, $-h \in \mathcal{H}$ and there exist $1 \le i \ne j \le n$ such that $y_i = h(x_i)$ and $y_j \ne h(x_j)$. Let $\mathcal{F}$ be the set of all linear combinations of elements of $\mathcal{H}$:

$$\mathcal{F} = \left\{ \sum_{j=1}^{M} \theta_j h_j \,;\, \theta \in \mathbb{R}^M \right\} .$$

Consider the following algorithm. Set $\hat{f}_0 = 0$ and for all $1 \le m \le M$,

$$\hat{f}_m = \hat{f}_{m-1} + \beta_m h_{j_m} \quad \text{where} \quad (\beta_m, h_{j_m}) = \underset{h\in\mathcal{H}\,,\,\beta\in\mathbb{R}}{\operatorname{argmin}}\; n^{-1} \sum_{i=1}^{n} \exp\left\{-y_i \left(\hat{f}_{m-1}(x_i) + \beta h(x_i)\right)\right\} .$$

1. Choosing $\omega_i^m = n^{-1} \exp\{-y_i \hat{f}_{m-1}(x_i)\}$, show that

$$n^{-1} \sum_{i=1}^{n} \exp\left\{-y_i \left(\hat{f}_{m-1}(x_i) + \beta h(x_i)\right)\right\} = \left(\mathrm{e}^{\beta} - \mathrm{e}^{-\beta}\right) \sum_{i=1}^{n} \omega_i^m \mathbb{1}_{h(x_i)\ne y_i} + \mathrm{e}^{-\beta} \sum_{i=1}^{n} \omega_i^m .$$

2. For all $1 \le m \le M$ and $h \in \mathcal{H}$, define

$$\mathrm{err}_m(h) = \frac{\sum_{i=1}^{n} \omega_i^m \mathbb{1}_{h(x_i)\ne y_i}}{\sum_{i=1}^{n} \omega_i^m} .$$

   Prove that

$$h_{j_m} = \underset{h\in\mathcal{H}}{\operatorname{argmin}}\; \mathrm{err}_m(h) \quad \text{and} \quad \beta_m = \frac{1}{2} \log\left(\frac{1 - \mathrm{err}_m(h_{j_m})}{\mathrm{err}_m(h_{j_m})}\right) .$$

3. Propose an algorithm to compute $\hat{f}_M$.

*** Hands on 6 Classification and fairness on the *adult* data set

See attached notebook.

Hands-on Session 8: SVM, RKHS

Friday 12th March, 2020

*** Exercise 11 SVM

Reminder on KKT conditions

Let $f, g_1, \ldots, g_n$ be $\mathcal{C}^1$ convex functions and define

$$\hat{x} = \underset{g_i(x) \leq 0}{\arg\min} f(x) .$$

Karush-Kuhn-Tucker necessary (& sufficient) conditions:
Define $L(x, \lambda) = f(x) + \sum_{i=1}^{n} \lambda_i g_i(x)$. Then, there exists $\hat{\lambda}$ such that

1. $\nabla_x L(\hat{x}, \hat{\lambda}) = 0$;
2. $\hat{\lambda}_i g_i(\hat{x}) = 0$ for $i = 1, \ldots, n$;
3. $g_i(\hat{x}) \leq 0$ for $i = 1, \ldots, n$;
4. $\hat{\lambda}_i \geq 0$ for $i = 1, \ldots, n$.

Strong duality: in addition $\hat{\lambda} = \underset{\lambda \geq 0}{\operatorname{argsup}} \inf_x L(x, \lambda)$.

For any $w \in \mathbb{R}^p$, define the linear function $f_w(x) = \langle w, x \rangle$ from $\mathbb{R}^p$ to $\mathbb{R}$. For a given $R > 0$, we consider the set of linear functions $\mathcal{F} = \{f_w : \|w\| \leq R\}$. The aim of this exercise is to investigate the classifier $\widehat{h}_{\varphi,\mathcal{F}}(x) = \mathsf{sign}(\widehat{f}_{\varphi,\mathcal{F}}(x))$ where $\widehat{f}_{\varphi,\mathcal{F}}$ is solution to the convex optimisation problem

$$\widehat{f}_{\varphi,\mathcal{F}} = \underset{f \in \mathcal{F}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \varphi(-y_i f(x_i)) ,$$

with $\varphi(x) = (1 + x)_+$ the *hinge* loss. The Lagrangian version of this minimization problem is

$$\widehat{f}_{\varphi,\mathcal{F}} = \underset{f_w : w \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} (1 - y_i f_w(x_i))_+ + \lambda \|w\|^2 \right\} ,$$

for some $\lambda > 0$.

1. Prove that $\widehat{f}_{\varphi,\mathcal{F}} = f_{\widehat{w}}$ where $\widehat{w}$ belongs to $V = \mathsf{Span}\{x_i : i = 1, \ldots, n\}$.
2. Prove that $\widehat{w} = \sum_{j=1}^{n} \widehat{\beta}_j x_j$ where $\widehat{\beta} = [\widehat{\beta}_1, \ldots, \widehat{\beta}_n]^T$ is solution to

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} (1 - y_i (K\beta)_i)_+ + \lambda \beta^T K \beta \right\} ,$$

with K the Gram matrix $K = [\langle x_i, x_j \rangle]_{1 \leq i,j \leq n}$.

3. Check that this minimization problem is equivalent to

$$\widehat{\beta} = \underset{\substack{\beta,\xi \in \mathbb{R}^n \text{ such that} \\ y_i(K\beta)_i \geq 1-\xi_i \\ \xi_i \geq 0}}{\arg\min} \left\{ \frac{1}{n}\sum_{i=1}^{n} \xi_i + \lambda \beta^T K \beta \right\}.$$

4. From the KKT conditions, check that $\widehat{\beta}_i = y_i \widehat{\alpha}_i/(2\lambda)$, for $i = 1, \ldots, n$ with $\widehat{\alpha}_i$ fulfilling $\min(\widehat{\alpha}_i, y_i(K\widehat{\beta})_i - (1-\widehat{\xi}_i)) = 0$ et $\min(1/n - \widehat{\alpha}_i, \widehat{\xi}_i) = 0$.

5. Prove the following properties
   - if $y_i \widehat{f}_{\varphi,\mathcal{F}}(x_i) > 1$ then $\widehat{\beta}_i = 0$;
   - if $y_i \widehat{f}_{\varphi,\mathcal{F}}(x_i) < 1$ then $\widehat{\beta}_i = y_i/(2\lambda n)$;
   - if $y_i \widehat{f}_{\varphi,\mathcal{F}}(x_i) = 1$ then $0 \leq \widehat{\beta}_i y_i \leq 1/(2\lambda n)$.

6. Give a geometric interpretation of this result.

7. From the strong duality, prove that $\widehat{\alpha}_i$ is solution to the dual problem

$$\widehat{\alpha} = \underset{0 \leq \alpha_i \leq 1/n}{\text{argmax}} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{4\lambda} \sum_{i,j=1}^{n} K_{i,j} y_i y_j \alpha_i \alpha_j \right\}.$$

** Exercise 12 RKHS

True or False? Either provide a proof (when true) or an explicit counterexample (when false)

1. If k_1 and k_2 are both positive semidefinite (PSD) kernel functions on $\mathcal{X} \times \mathcal{X}$, then $\lambda k_1 + \mu k_2$ is a PSD kernel function for all $\lambda, \mu > 0$.

2. Any Symmetric function k that is element-wise non-negative is a PSD kernel function.

3. If k_1 and k_2 are both positive semidefinite (PSD) kernel functions on $\mathcal{X} \times \mathcal{X}$, then $k(x,y) = k_1(x,y)k_2(x,y)$ is also a PSD kernel function.

4. Given a probability space with events $\mathcal{E}$ and probability law $\mathbb{P}$, the function $k : \mathcal{E} \times \mathcal{E} \to \mathbb{R}$ defined by $k(A,B) = \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)$ is a PSD kernel function.

5. Given a finite set $\mathcal{E}$, let $\mathcal{P}(\mathcal{E})$ denote the set of all subsets of $\mathcal{E}$. If $k : \mathcal{E} \times \mathcal{E} \to \mathrm{R}$ is a PSD kernel function, then

$$\bar{k}(A,B) = \sum_{x \in A,\ y \in B} k(x,y)$$

   is a PSD kernel function on $\mathcal{P}(\mathcal{E}) \times \mathcal{P}(\mathcal{E})$.

** Hands on 7 SVR on Time Series

See attached notebook.

Hands-on Session 9: Neural Networks

Friday 19th March, 2020

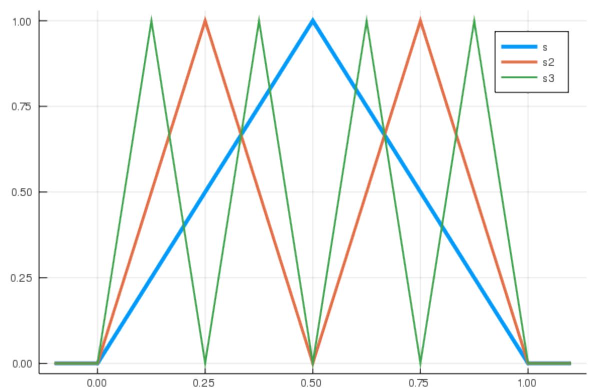
*** Exercise 13 The Expressive Power of Depth in Neural Networks

In this exercise, we consider ReLU networks, that is neural networks (with biases) whose transfer function is the the ReLU $r : \mathbb{R} \to \mathbb{R}$ defined by $r(x) = \max(x, 0)$.

1. Let $g : \mathbb{R} \to \mathbb{R}$ be constant outside of an interval $[0, R]$ and L-Lipschitz. Let $\epsilon > 0$ and $m = \lceil RL/\epsilon \rceil$. Show that the the piecewise linear function f coinciding with g at points $x_i = i\epsilon/L$, $i \in \{0, \ldots, m\}$, linear between x_i and x_{i+1}, and constant outside of $[0, x_m]$, is such that $\|f - g\|_\infty \leq \epsilon$.
2. If $\epsilon > RL$, find a very simple ReLU network f such that $\|f - g\|_\infty \leq \epsilon$.
3. If $\epsilon \leq RL$, show that the approximation f of Question 1 is implementable as a depth-2 ReLU network with linear output of width at most $m + 1 \leq 3RL/\epsilon$ and weights at most equal to $\max\left(2L, \|g\|_\infty\right)$.
4. Using the approximation of the previous question, how many neurons are required to approximate function function $x \mapsto x^2$ on $[0, 1]$ uniformly with an error at most equal to $\epsilon > 0$?
5. Let

$$s(x) = \begin{cases} 2x & \text{if } 0 \leq x \leq \frac{1}{2} \\ 2 - 2x & \text{if } \frac{1}{2} \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$
$$= 2r(x) - 4r\left(x - \frac{1}{2}\right) + 2r(x - 1) \,,$$

and for all $m \geq 1$ let $s_m = \underbrace{s \circ \cdots \circ s}_{m \text{ times}}$.



Plot (simple) ReLU networks implementing respectively s, s_2 and s_3.

6. Show that for all $m \geq 1$, all $k \in \left\{0, \ldots, 2^{m-1} - 1\right\}$ and all $t \in [0, 1]$,

$$s_m\left(\frac{k + t}{2^{m-1}}\right) = s(t)$$

7. Let $g(x) = x^2$, and for $m \geq 0$ let $g_m(x)$ be such that for all $k \in \left\{0, \ldots, 2^m\right\}$:
   - $g_m\left(\frac{k}{2^m}\right) = g\left(\frac{k}{2^m}\right)$,
   - g_m is linear on $\left[\frac{k}{2^m}, \frac{k+1}{2^m}\right]$.

   Show that for all $k \in \left\{0, \ldots, 2^m - 1\right\}$ and all $t \in [0, 1]$,

$$g_m\left(\frac{k + t}{2^m}\right) - g\left(\frac{k + t}{2^m}\right) = \frac{t(1 - t)}{4^m} \,.$$

8. Show that $\|g - g_m\|_\infty = \frac{1}{4^{m+1}}$ and for all $m \geq 2$,

$$g_m = g_{m-1} - \frac{1}{4^m}\, s_m = id - \sum_{j=1}^{m} \frac{1}{4^j}\, s_j$$

9. Deduce from the previous question (and plot) a neural network uniformly approximating g on $[0, 1]$ with a maximal error of $\epsilon > 0$.
10. Compare the networks of questions 4 and 9.

** Hands on 8 Experimenting Deep Learning

See attached notebook.

Hands-on Session 10: Reinforcement Learning

Friday 19th March, 2020

** Exercise 14 Bellman's Transition Operator

1. Show that Bellman's Transition Operator $T_\pi : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ defined by

$$T_\pi(V) = \bar{R}_\pi + \gamma K_\pi V$$

   is **affine**, **isotonic** ($U \leq V \implies T_\pi U \leq T_\pi V$) and **$\gamma$-contractant**: $\forall U, V \in \mathbb{R}^{\mathcal{S}}, \|T_\pi U - T_\pi V\|_\infty \leq \gamma \|U - V\|_\infty$
2. Show that T_π **has a unique fixed point equal to** V_π and that

$$\forall V_0 \in \mathbb{R}^{\mathcal{S}}, \;\; T_\pi^n V_0 \underset{n \to \infty}{\to} V_\pi \; .$$

**** Exercise 15 Bellman's Optimality Operator

Show that Bellman's Optimality Operator $T_* : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ defined by

$$\big(T_*(V)\big)_s = \max_{a \in \mathcal{A}} \left\{ \bar{r}(s,a) + \gamma \sum_{s' \in \mathcal{S}} k(s'|s,a) V_{s'} \right\} \; .$$

is **isotonic** and **γ-contractant**. Besides, for every policy π, $T_\pi \leq T_*$ in the sense that $\forall U \in \mathbb{R}^{\mathcal{S}}, T_\pi U \leq T_* U$.

*** Exercise 16 Policy Improvement Lemma

Prove that for any policy π, any greedy policy π' wrt V_π improves on π: $V_{\pi'} \geq V_\pi$.

*** Exercise 17 Bellman's Optimality Theorem

Prove that V_*, the unique fixed point of Bellman's optimality operator T_*, is the optimal value function:

$$\forall s \in \mathcal{S}, V_*(s) = \max_\pi V_\pi(s)$$

and any policy π such that $T_\pi V_* = V_*$ is optimal.

** Exercise 18 Correctness of the Value Iteration algorithm

Prove that the Value Iteration algorithm returns a value vector V such that $\|V - V_*\|_\infty \leq \epsilon$ using at most $\frac{\log \frac{M}{(1-\gamma)\epsilon}}{1-\gamma}$ iterations, where $M = \|T_* V_0 - V_0\|_\infty$.

*** Exercise 19 Policy Improvement Lemma: Q-table form

Prove that for any two policies π and π',

$$\Big[\forall s \in \mathcal{S}, Q_\pi\big(s, \pi'(s)\big) \geq Q_\pi\big(s, \pi(s)\big)\Big] \implies \Big[\forall s \in \mathcal{S}, V_{\pi'}(s) \geq V_\pi(s)\Big]$$

Furthermore, if one of the inequalities in the LHS is strict, then at least one of the inequalities in the RHS is strict

** Exercise 20 Bellman's Optimality Condition: Q-table formulation

Prove that a policy π is optimal if and only if

$$\forall s \in \mathcal{S},\ \pi(s) \in \operatorname*{argmax}_{a \in \mathcal{A}} Q_\pi(s, a)$$

** Hands on 9 Retail Shop Management

See attached notebook.