MAP569 Machine Learning II

PC9: Old exams, and other revisions

Instructions:

Every answer should be explained.

You don't need to answer all the questions to have a very good grade.

1 (Exam 2019) Problem - Reweighted Learning

In this problem, we study a generic machine learning scheme in which one observe some independent couples (X_i, Y_i) and try to find the best predictor $f_\theta(\tilde{X})$ according to a given loss $\ell(\tilde{Y}, f(\tilde{X})$ and a distribution Q for $(\tilde{X}, \tilde{Y})$ that may be different from the ones of the (X_i, Y_i):

$$\mathbb{E}\left[\ell(\tilde{Y}, f(\tilde{X}))\right]$$

We will assume that we have an algorithm that is able to minimize a weighted loss

$$\frac{1}{n}\sum_{i=1}^{n} w_i \ell'(Y_i, f_\theta(X_i))$$

for a loss that is related to ℓ but not necessarily equal.

1.1 Weighted loss

Assume for that $\ell(Y, f_\theta(X)) = w(X, Y)\ell'(Y, f_\theta(X))$ and that the (X_i, Y_i) are i.i.d. of law Q.

1. Justify the choice of $w_i = w(X_i, Y_i)$ in the empirical loss if our goal is to miminize $\mathbb{E}\left[\ell(Y, f_\theta(X))\right]$.
2. Assume we have a weighted least square algorithm, verify that one can deal with a relative least square loss

   $$\frac{(Y - f)^2}{Y^2 + \epsilon}$$

   but not with a relative least square loss

   $$\frac{2(Y - f)^2}{Y^2 + f^2}$$

3. Prove that, in the binary classification setting, starting from the 0/1 loss, $\ell'(Y, f) = 0$ if $Y = f$ and $\ell'(Y, f) = 1$ otherwise, one can find the minimizer for any choice of a binary loss $\ell(Y, f)$ defined by its four values.
4. Can we extend this result to any loss in a multiclass classification setting?

1.2 Importance Sampling

1. Assume that (X, Y) follows a law P with density dP with respect to a measure $d\lambda$, while $(\tilde{X}, \tilde{Y})$ follows a law Q with density dQ with respect to $d\lambda$. Prove that for any measurable function h

$$\mathbb{E}\left[h(\tilde{X}, \tilde{Y})\right] = \mathbb{E}\left[\frac{dQ(X, Y)}{dP(X, Y)} h(X, Y)\right]$$

as soon as $dP(X, Y) = 0 \Rightarrow dQ(X, Y) = 0$.

2. Prove that this formula involves only $dQ(X)/dP(X)$ if $P(Y|X) \sim Q(Y|X)$.

3. Assume that the observed (X_i, Y_i) are independent and such that for any i we assume that $(X_i, Y_i) \sim P_i$ while we are interested in an expected loss $\ell'(\tilde{Y}, f(\tilde{X}))$ with respect to $(\tilde{X}, \tilde{Y}) \sim Q$, how to choose the weight w_i so that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} w_i \ell'(Y_i, f_\theta(X_i))\right] = \mathbb{E}\left[\ell'(\tilde{Y}, f_\theta(\tilde{X}))\right]$$

1.3 Stratification, Reweighting and Unbalanced Dataset

We consider the following stratified sampling scenario in a multiclass classification setting.

- we know the probabilities $Q(\tilde{Y} = k)$ in the real world for all the K considered classes.
- the dataset is obtained class by class by sampling uniformly n_k samples in each of them.

1. Verify that in any class

$$\mathbb{E}\left[\ell'(\tilde{Y}, f_\theta(\tilde{X})\middle|\tilde{Y} = k\right]$$

can be estimated by a unweighted empirical loss.

2. Using the decomposition

$$\mathbb{E}\left[\ell'(\tilde{Y}, f_\theta(\tilde{X}))\right] = \sum_k Q\left(\tilde{Y} = k\right)\mathbb{E}\left[\ell'(\tilde{Y}, f_\theta(\tilde{X})\middle|\tilde{Y} = k\right],$$

propose a global weighting scheme to correct the sampling bias.

3. How to adapt this equality if we are interested in

$$\mathbb{E}\left[\ell(\tilde{Y}, f_\theta(\tilde{X}))\right]$$

with $\ell(\tilde{Y}, f) = C(\tilde{Y})\ell'(\tilde{Y}, f)$

4. How to use this formula in an unbalanced dataset setting in which
   - the proportions of the classes can be very different,
   - the proportions in the training dataset do not necessarily correspond to the one in the real world,
   - the cost of an error depends on the true class, i.e. $\ell(\tilde{Y}, f) = C(\tilde{Y})\ell'(\tilde{Y}, f)$?

Some complements on Stratification

Let $I(h) = \mathbb{E}[h(Z)] = \int h(z)f(z)dz$ where the integral is on $\mathbb{R}^d$ and f is some density function. Assume that there exists a partition of $\mathbb{R}^d$ into K regions, $D_1, \dots, D_K$. Write $\mu_i = \mathbb{E}[h(Z)|Z \in D_i]$ and $\sigma_i^2 = \mathrm{Var}[h(Z)|Z \in D_i]$. Assume that we know $\alpha_i = \mathrm{P}(Z \in D_i)$.

1. Propose an estimator $\tilde{S}_n$ of $I(h)$ that uses the α_i.
2. Give the expression of $\mathrm{Var}(\tilde{S}_n)$. Can we compare it to the rough estimator $S_n = n^{-1}\sum_{i=1}^n h(Z_i)$ of $I(h)$ where (Z_i) are iid according to the common density f?
3. In the case of proportional allocation ($n_i/n = \alpha_i$), show that $\mathrm{Var}(\tilde{S}_n) \leq \mathrm{Var}(S_n)$.
4. What is the optimal allocation? Any comments?

(PC8) Expectation Maximization algorithm

In the case where we are interested in estimating unknown parameters $\theta \in \mathbb{R}^m$ characterizing a model with missing data, the Expectation Maximization (EM) algorithm (Dempster et al. 1977) can be used when the joint distribution of the missing data X and the observed data Y is explicit. For all $\theta \in \mathbb{R}^m$, let p_θ be the probability density function of (X, Y) when the model is parameterized by θ with respect to a given reference measure μ. The EM algorithm aims at computing iteratively an approximation of the maximum likelihood estimator which maximizes the observed data loglikelihood:

$$\ell(\theta; Y) = \log p_\theta(Y) = \log \int f_\theta(x, Y)\mu(\mathrm{d}x)\,.$$

As this quantity cannot be computed explicitly in general cases, the EM algorithm finds the maximum likelihood estimator by iteratively maximizing the expected complete data loglikelihood.

1. Recall the two steps of an iteration of the EM algorithm.
2. Prove that the loglikelihood monotonically increases along EM iterations.

Let M_n^+ the space of real-valued $n \times n$ symmetric positive matrices. We first show that the function $X \mapsto \log \det X$ is concave on M_n^+.

3. Let $X, Y \in M_n^+$ and $\lambda \in [0, 1]$. Since $X^{-1/2}YX^{-1/2} \in M_n^+$, it is diagonalisable in some orthonormal basis and write $\mu_1, \dots, \mu_n$ the (possibly repeated) entries of the diagonal. Show that
$$\log\det\{(1-\lambda)X + \lambda Y\} \geq \log \det X + \lambda \sum_{i=1}^n \log(\mu_i)$$
4. Conclude.

In the following, $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_n)$ where $\{(X_i, Y_i)\}_{1 \leqslant i \leqslant n}$ are i.i.d. in $\{-1, 1\} \times \mathbb{R}^d$. For $k \in \{-1, 1\}$, write $\pi_k = \mathbb{P}(X_1 = k)$. Assume that, conditionally on the event $\{X_1 = k\}$, Y_1 has a Gaussian distribution with mean $\mu_k \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. In this case, the parameter $\theta = (\pi_1, \mu_1, \mu_{-1}, \Sigma)$ belongs to the set $\Theta = [0, 1] \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{d\times d}$.

5. Write the complete data loglikelihood.
6. Let $\theta^{(t)}$ be the current parameter estimate. Compute $\theta \mapsto Q(\theta, \theta^{(t)})$ (tips: use $\omega_t^i = \mathbb{P}_{\theta^{(t)}}(X_i = 1|Y_i)$)
7. Compute $\theta^{(t+1)}$.

(PC 7) RKHS

Let $(X_i)_{1\leq i\leq n}$ be n observations in a general space $\mathcal{X}$ and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a positive kernel. $\mathcal{W}$ denotes the Reproducing Kernel Hilbert Space associated with k and for all $x \in \mathcal{X}$, $\phi(x)$ denotes the function $\phi(x) : y \to k(x, y)$. The aim is now to perform a PCA on $(\phi(X_1), \ldots, \phi(X_n))$. It is assumed that

$$\sum_{i=1}^{n} \phi(X_i) = 0\,.$$

Define

$$\mathbf{K} = (k(X_i, X_j))_{1\leqslant i,j\leqslant n}\ .$$

1. Prove that

$$f_1 = \underset{f\in\mathcal{W}\,;\,\|f\|_{\mathcal{W}}=1}{\operatorname{argmax}} \sum_{i=1}^{n} \langle \phi(X_i), f\rangle_{\mathcal{W}}^2$$

   may be written

$$f_1 = \sum_{i=1}^{n} \alpha_1(i)\phi(X_i)\ , \quad \text{where} \quad \alpha_1 = \underset{\alpha\in\mathbb{R}^n\,;\,\alpha^T\mathbf{K}\alpha=1}{\operatorname{argmax}} \alpha^T\mathbf{K}^2\alpha\ .$$

2. Prove that $\alpha_1 = \lambda_1^{-1/2} b_1$ where b_1 is the unit eigenvector associated with the largest eigenvalue λ_1 of $\mathbf{K}$.
3. Following the same steps, f_j may be written $f_j = \sum_{i=1}^{n} \alpha_j(i)\phi(x_i)$ with $\alpha_j = \lambda_j^{-1/2} b_j$. Write $H_d = \mathrm{span}\{f_1, \ldots, f_d\}$. Prove that

$$\pi_{H_d}(\phi(x_i)) = \sum_{j=1}^{d} \lambda_j \alpha_j(i) f_j\ .$$

2 Short Questions

We expect an answer of no more than 2-3 lines for any of those questions.

1. Why is the training error an optimistic estimate of the generalization error?
2. What are the support vectors in a SVM?
3. What is the principle of the back-prop algorithm?
4. What is a gradient boosting algorithm?
5. Why is the k-means clustering algorithm easy to distribute?