Statistics review

Semaine de pré-rentrée du master MVA

In the following list of exercises, we use the notation $\boldsymbol{X}^\top$ (resp. $\mathbf{x}^\top$) to denote the transpose of the matrix $\boldsymbol{X}$ (resp. the vector $\mathbf{x}$).

Multinomial random variables

The vector $(N_1, \dots, N_K)$ is said to follow a multinomial distribution $\mathcal{M}(\pi_1, \dots, \pi; n)$ if and only for any non-negative integers $n_1, \dots, n_k$ we have

$$\mathbb{P}(N_1 = n_1, \dots, N_K = n_K) = \binom{n}{n_1, \dots, n_K} \prod_{k=1}^{K} \pi_k^{n_k} \, 1_{\{n_1 + \dots + n_K = n\}}.$$

Of particular interest is the special case where $n = 1$ which is quite convenient to encode probability distributions with finite discrete support.

1. Show that if $Z = (Z_1, \dots, Z_K) \sim \mathcal{M}(\pi_1, \dots, \pi; 1)$ then Z is a binary indicator vector with $\mathbb{P}(Z_k = 1) = \pi_k$.

2. If $Z^{(1)}, \dots, Z^{(n)}$ is an i.i.d. sample from $\mathcal{M}(\pi_1, \dots, \pi; 1)$ then defining $N_k = \sum_{i=1}^n Z_k^{(i)}$ for all k, show that $N := (N_1, \dots, N_K)$ follows the distribution $\mathcal{M}(\pi_1, \dots, \pi; n)$.

Method of moments *vs* maximum likelihood estimation

1. ($\star$) **Uniform distributions.** In this exercise the Beta distribution will be useful. Remember that the Beta-distribution with parameters $\alpha, \beta > 0$ is the distribution on the interval $[0, 1]$ with density $p_{\alpha,\beta}(x) \propto x^{\alpha-1}(1-x)^{\beta-1}$. Its mean and variance are respectively equal to $\alpha/(\alpha+\beta)$ and $\alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$. We consider the *statistical model* consisting of the uniform distributions on the interval $[0, \theta]$ for some $\theta > 0$ and undertake to estimate θ from a sample $X_1, \dots, X_n$ drawn from such a distribution. Since, $\mathbb{E}[X_i] = \theta/2$, one candidate estimator is the moment estimator $\hat{\theta}_{MO} = \frac{2}{n} \sum_{i=1}^n X_i$.

   (a) Show that the maximum likelihood estimator $\hat{\theta}_{MLE}$ exists and is unique and compute it.

   (b) Show that $\hat{\theta}_{MLE}$ follows a Beta distribution. What are the values of the parameters of this Beta?

   (c) Deduce from the previous question the variance and the bias of the estimator.

   (d) What is the variance of the moment estimator?

   (e) We consider the MSE $\mathbb{E}[(\theta - \hat{\theta})^2]$ as a measure of performance of the estimator. Compare the MSE for both estimators. Which estimator should be preferred?

Computation of maximum likelihood estimators

1. Let $x_1, \ldots, x_n$ be an i.i.d. sample from a Bernoulli distribution with parameter θ.

   (a) Use convexity arguments to decide if the MLE exists and whether it is unique.

   (b) Compute the MLE

2. Consider the random binary indicator vector $Z = (Z_1, \ldots, Z_K)$ following the multinomial distribution $\mathcal{M}(\pi_1, \ldots, \pi_K; 1)$, in other words such that $\mathbb{P}(Z_k = 1) = \pi_k$. Let $z^{(1)}, \ldots, z^{(n)}$ be a sample of such multinomial vectors.

   (a) Let $N_k = \sum_{i=1}^n Z_k^{(i)}$. Show that $(N_1, \ldots, N_K)$ is a *sufficient statistic* for the sample and express the likelihood as a function of $(N_1, \ldots, N_K)$.

   (b) Show that the MLE is the solution of a constrained convex optimization problem (in particular argue that the MLE exists and is unique).

   (c) Construct the associated Lagrangian and derive the MLE.

3. ($\star$) Consider the multivariate Gaussian r.v. $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$. Assume that $\boldsymbol{\Sigma}$ is positive definite (and not only semi-definite), so that the Gaussian distribution admits a density with respect to the Lebesgue measure of the form:

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\Big\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \Big\}.$$

   ($*$) Let $\boldsymbol{A} \in \mathbb{R}^{d \times d}$, consider the functions $f : \boldsymbol{\mu} \mapsto \boldsymbol{u}^\top \boldsymbol{\mu}$ and $g : \boldsymbol{\mu} \mapsto \boldsymbol{\mu}^\top \boldsymbol{A} \boldsymbol{\mu}$. Compute the differential [1] $df_{\boldsymbol{\mu}} : \mathbb{R}^d \to \mathbb{R}$ of f at $\boldsymbol{\mu}$ and deduce the form of the gradient $\nabla f(\boldsymbol{\mu})$. Do the same for $dg_{\boldsymbol{\mu}}$ and $\nabla g(\boldsymbol{\mu})$. How are these simplified if $\boldsymbol{A}$ is a symmetric matrix?

   We now assume that we have an i.i.d. sample $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ from the above-mentionned multivariate Gaussian distribution.

   (a) Show that if $\boldsymbol{\Sigma}$ is fixed, the MLE estimator for $\boldsymbol{\mu}$ is the minimum of a quadratic form and that it does not depend on $\boldsymbol{\Sigma}$. What is the MLE equal to?

   (b) Assuming now that $\boldsymbol{\mu}$ is fixed show that the statistic

$$\widehat{\boldsymbol{\Sigma}} := \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top,$$

   is a sufficient statistic for $\boldsymbol{\Sigma}$ and express the log-likelihood as a function of $\widehat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Lambda} := \boldsymbol{\Sigma}^{-1}$.

   (c) Given the Frobenius inner product $\langle \cdot, \cdot \rangle_F$ between matrices, with $\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F = \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{B})$, remember that for a differentiable real valued matrix function $f : \mathbb{R}^{p \times d} \to \mathbb{R}$, the gradient of f at $\boldsymbol{A}$ is defined as the matrix $\boldsymbol{\nabla} f(\boldsymbol{A})$ such that for all $\boldsymbol{H} \in \mathbb{R}^{p \times d}$, we have

$$f(\boldsymbol{A} + \boldsymbol{H}) - f(\boldsymbol{A}) = \langle \boldsymbol{\nabla} f(\boldsymbol{A}), \boldsymbol{H} \rangle_F + o(\|\boldsymbol{H}\|_F).$$

   (d) What is the gradient of the function $\boldsymbol{B} \mapsto \langle \boldsymbol{A}, \boldsymbol{B} \rangle_F$?

   (e) If $\boldsymbol{I}$ is the identity matrix and if $|\cdot|$ denotes the determinant of a matrix, show that the differential of the log-determinant [2] at the identity is the trace, that is $d|\cdot|_{\boldsymbol{I}}(\boldsymbol{H}) = \mathrm{tr}(\boldsymbol{H})$. Deduce from this the form of the differential of the log-determinant restricted to the set of symmetric matrices at a matrix $\boldsymbol{A}$ (which we also assume symmetric), and the value of the gradient of the log-determinant at $\boldsymbol{A}$.

[1] Remember that the differential of a real valued differentiable function f at $\mathbf{x} \in \mathbb{R}^m$ is the linear form $df_{\mathbf{x}} : \mathbb{R}^m \mapsto \mathbb{R}$ such that for all $\boldsymbol{h} \in \mathbb{R}^m$, we have $f(\mathbf{x} + \boldsymbol{h}) - f(\mathbf{x}) = df_{\mathbf{x}}(\boldsymbol{h}) + o(\|\boldsymbol{h}\|)$.

[2] the logarithm of the determinant

(f) Compute the gradient of the log-likelihood of a sample with $\boldsymbol{\mu}$ fixed and compute the maximum likelihood estimator for $\boldsymbol{\Lambda}$. Deduce the MLE for $\boldsymbol{\Sigma}$?

(g) Consider now the computation of the joint MLE for the pair $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$? What changes? What happens if $\widehat{\boldsymbol{\Sigma}}' := \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$ is not invertible?

Ridge regression and PCA

1. **Moore-Penrose pseudo-inverse.** The Moore-Penrose pseudo-inverse of a rectangular matrix $\boldsymbol{S} \in \mathbb{R}^{n\times p}$ that has non-zeros entries only on the diagonal is the matrix $\boldsymbol{S}^- \in \mathbb{R}^{p\times n}$ which is such that $(\boldsymbol{S}^-)^\top$ is the matrix whose zero entries are the same as those of $\boldsymbol{S}$ and with non-zero diagonal entries the inverses of the non-zero diagonal entries of $\boldsymbol{S}$ (note that some of the entries of the diagonal can be equal to zero, in which case they are zero in both matrices). More generally the Moore-Penrose pseudo-inverse of a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ with singular value decomposition $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ (where $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{V} \in \mathbb{R}^{p\times p}$ are orthogonal matrices, and $\boldsymbol{S} \in \mathbb{R}^{n\times p}$ is a rectangular diagonal matrix containing the singular values of $\boldsymbol{X}$) is the matrix $\boldsymbol{X}^- = \boldsymbol{V}\boldsymbol{S}^-\boldsymbol{U}^\top$. The Moore-Penrose pseudo-inverse is often denoted $\boldsymbol{X}^\dagger$.

   (a) Show that for any matrix $\boldsymbol{X}$, we have $\boldsymbol{X}\boldsymbol{X}^-\boldsymbol{X} = \boldsymbol{X}$ and $\boldsymbol{X}^-\boldsymbol{X}\boldsymbol{X}^- = \boldsymbol{X}^-$

   (b) Show that $(\boldsymbol{X}^\top\boldsymbol{X})^- = \boldsymbol{X}^-(\boldsymbol{X}^-)^\top$ and that $(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top = \boldsymbol{X}^-$.

   (c) Show that $\boldsymbol{w}_{\mathrm{PI}} := \boldsymbol{X}^-\boldsymbol{y}$ is a solution to the *normal equation* $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^\top\boldsymbol{y}$.

   (d) Show that it is the solution of the *normal equation* with minimal Euclidean norm.

2. **Review of PCA.** For data $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ living in $\mathbb{R}^p$, PCA defines a sequence of mutually orthogonal vectors called *principal directions* (or *principal factors*) $\boldsymbol{v}_1, \boldsymbol{v}_2$, etc $\in \mathbb{R}^p$ that such that the projection of the data on these directions have maximal variance. These directions are used to define new variables called *principal variables* or *principal components*. In particular the value of the jth principal component for the ith datapoint $\mathbf{x}^(i)$ is $c_{ij} := \langle \mathbf{x}^{(i)}, \boldsymbol{v}_j \rangle$. The jth principal variable is represented in the dataset by the vector $\boldsymbol{c}_j = \boldsymbol{X}\boldsymbol{v}_j \in \mathbb{R}^n$. We will assume in this exercise that $\boldsymbol{X}$ is centered (in the sense that each column has zero mean.)

   (a) Show that the $(\boldsymbol{v}_j)_j$ are the right singular vectors of the design matrix.

   (b) Show that the $(\boldsymbol{c}_j)_j$ can be retrieved from the Gram matrix $\boldsymbol{K} := \boldsymbol{X}\boldsymbol{X}^\top$ only.

   (c) Show that the empirical standard deviation of the entries in $\boldsymbol{c}_j$ is equal to the jth singular value σ_j of $\boldsymbol{X}$.

   (d) Show that $\sigma_j^{-1}\boldsymbol{c}_j$ is $\boldsymbol{u}_j$ the jth left singular vector of $\boldsymbol{X}$.

3. ($\star$) **Ridge regression and PCA.** Karl and Andreï have to solve a linear regression problem and they are debating about which method to use. Given that the amount of data is not so large as compared to the number of covariates[3] they are concerned about overfitting. They have a sample of training data consisting of a number of pairs $(\mathbf{x}_0^{(i)}, y^{(i)})$ with $\mathbf{x}_0^{(i)} \in \mathbb{R}^p$ and $y^{(i)} \in \mathbb{R}$. They first compute $\bar{\mathbf{x}}_0$ the empirical average of the vectors $(\mathbf{x}_0^{(i)})_i$, compute the centered vectors $\mathbf{x}^{(i)} = \mathbf{x}_0^{(i)} - \bar{\mathbf{x}}_0$ and store them in a *design matrix* $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ where, as usual, each row corresponds to an observation $\mathbf{x}^{(i)}$ and each column to a covariate.

   Andreï is in favor of using *ridge regression*, which he learnt about in class, while Karl has another idea: Karl has heard about parsimony and would prefer to reduce the number of parameters of the regression, so that regularization is no longer necessary; to be precise, he suggests to replace the set of initial covariates by the k first principal components of the point cloud and to construct a prediction model based on them.

[3] *covariates* or *variables* in statistics, *features* or *descriptors* in ML

The goal of the exercise is to compare the two methods. With the notation of the previous exercise, the idea of Karl is to project the training data on the principal directions $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$, to keep only the obtained principal components $\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_k$ and to use these as a new set of covariates to learn a regression function without regularization. Then in order to make new prediction, the new datapoint can be projected on the $(\boldsymbol{v}_j)$ and from these projections, a prediction can be made.

(a) Denoting $\boldsymbol{c}_{1:k}^{(i)} := (c_{i1}, \ldots, c_{ik})$, consider the estimate $\tilde{\boldsymbol{w}}$ for the parameter vector of the regression of the $y^{(i)}$s on the $\boldsymbol{c}_{1:k}^{(i)}$s. Show that $\tilde{\boldsymbol{w}}$ can be expressed as a function of $\boldsymbol{y} = (y^{(1)}, \ldots, y^{(n)})$, of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$, the k first left singular vectors of $\boldsymbol{X}$, and of $\sigma_1, \ldots, \sigma_k$, the k largest singular values of $\boldsymbol{X}$.

(b) Given a new datapoint $\mathbf{x}$ show how the linear function estimated from the principal components translate into an affine predictor on $\mathbf{x}$ of the form $\mathbf{x} \mapsto \boldsymbol{w}_{\text{PCA}}^\top(\mathbf{x} - \bar{\mathbf{x}}_0)$ where $\boldsymbol{w}_{\text{PCA}}$ is of the form $\sum_{j=1}^k \gamma_j \boldsymbol{v}_j$ where $\gamma_j \in \mathbb{R}$ depends only on $\boldsymbol{u}_j$, $\boldsymbol{y}$ and σ_j.

(c) Let $m = \text{rank}(\boldsymbol{X})$. Now, given a regularization coefficient λ, show that the predictor from ridge regression is of the form $\mathbf{x} \mapsto \boldsymbol{w}_{\text{R}}^\top(\mathbf{x} - \bar{\mathbf{x}}_0)$, where $\boldsymbol{w}_{\text{R}} \in \mathbb{R}^m$ is of the form $\sum_{j=1}^m \gamma_j'(\lambda)\boldsymbol{v}_j$ where $\gamma_j(\lambda) \in \mathbb{R}$ depends only on $\boldsymbol{u}_j$, $\boldsymbol{y}$, λ and σ_j.

(d) Compare $\boldsymbol{w}_{\text{PCA}}$ with $\boldsymbol{w}_{\text{R}}$, in what way are they similar? In particular, given the fact that the function $\psi : z \mapsto z^2/(\lambda + z^2)$ is an increasing function from 0 to 1 and that $\psi(\sqrt{\lambda}) = 0.5$, explain why keeping the k-first principal component is somewhat like choosing λ between σ_k^2 and σ_{k+1}^2.

(e) If you know the Moore-Penrose pseudo-inverse or if you have done the exercise on it (see that exercise for notations), show that denoting $\boldsymbol{w}_{\text{PI}} := \boldsymbol{X}^\top \boldsymbol{y}$, we have that $\boldsymbol{w}_{\text{PCA}}$ is the projection of $\boldsymbol{w}_{\text{PI}}$ on the k-first right singular vectors of $\boldsymbol{X}$ and that ridge regression produces some "soft-projection" of $\boldsymbol{w}_{\text{PI}}$ on these vectors. To which vector does the solution of ridge regression converge when $\lambda \to \infty$.

(f) Why are the two characters called Andreï and Karl?

Sufficient statistic

Given a random variable X (which is typically a sample $X = (X^{(1)}, \ldots, X^{(n)})$) drawn from a a distribution with density p_θ for some $\theta \in \Theta$, the statistic $T = T(X)$ is said to be a *sufficient statistic* (in french *statistique exhaustive*) if there exist functions f and h such that for any x

$$p_\theta(x) = h(x, T(x)) f(T(x); \theta).$$

- $(\star)$ Consider the point of a view of a Bayesian statistician who treats θ as a random variable. Assume that the joint distribution of (X, θ) has the density $p(x, \theta) = p(x|\theta)p(\theta)$ with $p(x|\theta) = p_\theta(x)$ and with $p(\theta)$ the prior distribution. Then show that T is a sufficient statistic if and only if θ and X are conditionally independent given T.

Bayesian estimation

(a) **Bayesian estimation of a multinomial.** Distributions in the Dirichlet family put mass only on the simplex $\triangle := \{\boldsymbol{u} \in \mathbb{R}_+^K \mid u_1 + \ldots + u_K = 1\}$ and they admit a density with respect to the Lebesgue measure on the simplex of the form

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1 + \ldots + \alpha_K)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \pi_k^{\alpha_k - 1},$$

where Γ is the usual Gamma-function defined as $\Gamma(z) = \int_{\mathbb{R}_+} t^{z-1}e^{-t}dt$ and where $\alpha_k > 0$ for all k. As a reminder, for all $z > 0$, the identity $\Gamma(z+1) = z\Gamma(z)$ holds and in particular for any $n \in \mathbb{N}_*$, $\Gamma(n) = n!$

We consider the Bayesian estimation of the parameters of a multinomial random variable. Assume that we have an i.i.d. sample $Z^{(1)}, \dots, Z^{(n)}$ from $\mathcal{M}(\pi_1, \dots, \pi; 1)$ (see the exercise on multinomial random variables for the notations) and that we consider a Dirichlet a priori distribution with hyperparameter $\boldsymbol{\alpha} \in \mathbb{R}_+^K$ for the vector of parameters $(\pi_1, \dots, \pi_K)$.

i. Show that the a posteriori distribution is also a Dirichlet distribution. What are its parameters?

ii. The most classical point estimator used to summarize the Bayesian a posterori given a sample $D_n = \{Z^{(1)}, \dots, Z^{(n)}\}$ is the posterior mean $\mathbb{E}[\theta|D_n] = \int_\Theta \theta\, p(\theta|D_n)\, d\theta$. Compute the posterior mean for the Multinomial-Dirichlet model.

iii. The Bernoulli (or binomial) distribution is the a particular case of a multinomial distribution with $K = 2$ in that case the Dirichlet distribution reduces to the Beta distribution. For the hyperparameters $\alpha_1 = \alpha_2 = 1$ compute the mean posterior estimate for the parameter π_1. This estimator is called the smooth Laplace estimate for the Bernoulli distribution. What are the advantages of this estimator?

(b) **Conjugate priors.** Given a model $\mathcal{P} = \{p(x|\theta) \mid \theta \in \Theta\}$ a conjugate prior family is collection of prior distributions $\Pi = \{p_\alpha(\theta) \mid \alpha \in A\}$ such that for all $p_\alpha \in \Pi$ we have $p(\theta|x;\alpha) := \int \frac{p(x|\theta)p(\theta;\alpha)}{p(x;\alpha)} d\theta$ satisfies $p(\theta|x;\alpha) \in \Pi$. What is the smallest canonical exponential family of distribution which is a conjugate prior family for the family of Bernoulli distributions, of Poisson distributions? What is the canonical exponential family of distribution which is a conjugate prior family for the family of Gaussian random variables with fixed known covariance matrix and unknown mean?

(c) ($\star$) **Bayesian estimation for the Gaussian.** Let $X \sim \mathcal{N}(\mu, \sigma^2)$ be a Gaussian random variable. Consider the problem of estimation of μ in the Bayesian sense, with an a priori distribution on μ of the form $\mu \sim \mathcal{N}(\mu_0, \tau^2)$. We will call μ_0 the prior mean.

i. Given an iid sample $X_1, \dots, X_n$ from the model, compute the posterior mean for μ.

ii. Show that the posterior mean $\hat{\mu}_{\text{PM}}$ is a convex combination of the MLE and the prior mean.

iii. Consider now the MAP estimator $\hat{\mu}_{\text{MAP}} := \arg\max p(X_1, \dots, X_n|\mu)p(\mu)$.

iv. Show that, in this case and if $\mu_0 = 0$, the MAP estimator can be viewed as a minimizer of the log-likelihood with some Tikhonov (or ridge) regularization.

v. Compute the MAP estimator. What is the relationship between the MAP estimator and the mean a posteriori? Do you expect this property to hold for other situations than the Gaussian distribution with a Gaussian prior on the mean?

vi. Show that among all Gaussian distributions with fixed variance σ^2 and some mean ν, the one that has the largest expected log-likelihood on the test random variable X' is the distribution whose mean minimizes what we will call the quadratic risk[4] $\mathcal{R}(\nu) = \mathbb{E}_{X'}[(\nu - X')^2]$.

vii. Show that if $\mathbb{E}[X'] = \mu$, we have the bias-variance decomposition $\mathcal{R}(\nu) = (\mu - \nu)^2 + \mathrm{Var}(X')$. What does this prove about the minimizer? Explain why we can focus on the *excess risk* $\mathcal{E}(\nu) = \mathcal{R}(\nu) - \mathcal{R}(\mu)$.

viii. Denote by $D_n = \{X_1, \dots, X_n\}$ the sample previously considered and denote by $\hat{\mu}$ an estimator of μ based on this sample. Show that the expected excess risk[5] $\mathbb{E}_{D_n}[\mathcal{E}(\hat{\mu})]$ can also be decomposed in terms of bias and variance.

[4] Note that the notion of risk introduced here is the notion of risk used in statistical learning theory and different from the notion of risk of an estimator associated with a contrast usually defined in classical statistics.

[5] Note that the quantity which is called the excess risk in statistical learning theory is in fact a contrast in the terminology of classical statistics and that the expected excess risk is what classical statisticians call *the risk*. It is important not to get confused by these different choices of terminology.

ix. Use this last bias-variance decomposition to compute the expected excess risk of $\hat{\mu}_{\mathrm{MLE}}$ and $\hat{\mu}_{PM}$. When does $\hat{\mu}_{PM}$ have smaller expected excess risk?

x. Given an estimator $\hat{\mu}$ of μ, we consider the Bayesian risk $\mathcal{R}_\pi = \mathbb{E}_{\mu\sim\pi}\mathbb{E}_{D_n}[\mathcal{E}(\hat{\mu})]$, where π denotes here the prior distribution assumed on μ namely the Gaussian distribution $\mathcal{N}(\mu_0, \tau^2)$. Note that since $\hat{\mu}$ is a function of the data D_n, which itself depends on μ, $\mathcal{R}_\pi$ is not a function of the estimator value $\hat{\mu}$, but characterizes the average behavior of the estimation procedure over the a priori distribution π. Explain why one can equivalently consider $\mathbb{E}_{\mu\sim\pi, D_n}[(\hat{\mu}-\mu)^2]$? Using the calculations of risk of $\hat{\mu}_{\mathrm{MLE}}$ and $\hat{\mu}_{\mathrm{PM}}$ done in the previous questions, compute the Bayesian risks $\mathcal{R}_\pi(\mathrm{MLE})$ and $\mathcal{R}_\pi(\mathrm{PM})$ for these two estimation procedures.

xi. Show that with the quantities computed, $\mathcal{R}_\pi(\mathrm{PM}) < \mathcal{R}_\pi(\mathrm{MLE})$.

xii. Using the expression of the Bayesian risk given above show that, this is not a coincidence and that in fact the Bayesian quadratic risk is minimized by the posterior mean estimator for any prior distribution π and for any likelihood.

Bregman divergence

The concept of Bregman divergence provides a generalization of the squared Euclidean distance which is quite relevant in statistics, optimization and machine learning. Given a continuously-differentiable strictly convex function F, called the *potential* function, and defined on a closed convex set of a Hilbert space, the associated Bregman divergence is defined as the function

$$D_F(p,q) = F(p) - [F(q) + \langle \nabla F(q), p-q \rangle].$$

(a) Show that if F is the squared Euclidean norm in $\mathbb{R}^d$, the associated divergence is the squared Euclidean norm.

(b) Consider two probability distributions $p = (p_i)_{1\leq i\leq d}$ and $q = (q_i)_{1\leq i\leq d}$ on a finite space. We define respectively the entropy $H(p)$ of the distribution p and the Kullback-Leibler divergence $KL(p,q)$ between the distributions p and q as

$$H(p) = -\sum_{i=1}^{q} p_i \log p_i \quad \text{and} \quad KL(p,q) = \sum_{i=1}^{q} p_i \log \frac{p_i}{q_i},$$

with the conventions $0/0 = 0$ and $0 \log 0 = 0$. Show that $KL(p,q)$ is the Bregman divergence $D_H(p,q)$ associated with the entropy.

(c) Let $\ell : (\mu, X) \mapsto \ell(\mu, X)$ be a *loss function*[6] and $\mathcal{R}(\mu) = \mathbb{E}[\ell(\mu, X)]$ the associated risk, where the expectation is taken w.r.t. the variable X. Denote by μ^* the minimizer of the risk, which is often called the *target parameter*, and consider the so-called *excess risk* $\mathcal{E}(\mu) := \mathcal{R}(\mu) - \mathcal{R}(\mu^*)$. Show that if the loss ℓ is strictly convex w.r.t. to its first argument and that $\mathcal{R}$ is differentiable, the *excess risk* can actually be interpreted as a Bregman divergence between μ and μ^*. What is the associated potential function ?

Area under the curve and Mann-Whitney U statistic

($\star$) The Mann-Whitney U-test also called Wilcoxon rank-sum test is a non parametric test to compare (non-paired) samples from two distributions. The null hypothesis is that the two samples are from the same distribution, and the alternative hypothesis is that the distribution differ. This test is

[6] A loss function is simply a function whose arguments are a parameter μ and an a random variable and which measure a certain discrepancy between them.

way to generalize two-sample t-test that rely on a Gaussian assumption. If the two samples are $D_X = \{x_1, \dots, x_n\}$ and $D_Y = \{y_1, \dots, y_m\}$ then we define the rank[7] of x_i as

$$r_i = \big|\big\{z \in D_X \cup D_Y \mid z \leq x_i\big\}\big|.$$

So for example if $n=3, m=2$ and $x_1=1, x_2=4, x_3=7$ and $y_1=2, y_2=5$ then we have $r_1=1, r_2=3, r_3=5$. The U statistic is defined as

$$U = \sum_{i=1}^{n} r_i - \frac{n(n+1)}{2}.$$

Under the null hypothesis the distribution of U does not depend on the distribution of the data, and its quantile can be computed to define a rejection region for the test.

Consider now a classifier that produces scores $s(x)$ for data to be classified in two classes, with the scores from class 0 that tend to be smaller than the scores for class 1. The classification rule used is that, if $s(x) > b$, then x is assigned to class 1 otherwise it is assigned to class 0. For any value of b, we define the recall or true positive rate rTP of the classifier as the fraction of true positives (or elements of class 1) that are predicted to be positive. We also define the false positive rate rFP of the classifier as the fraction of true negatives (or class 0) that are predicted to be positives by the classifier. The ROC curve is the curve that plots rFP as a function of rTP.

(a) Assuming that the fraction of positives in the testing set is π and that the testing set is very large what is the equation of the ROC curve for scores $s(x)$ that are produced at random following a continuous distribution that does not depend on the value of x? What is the area under the curve?

(b) To compare the performance of different classifiers, the area under the ROC curve called AUC is often computed. Assuming that there are no ties, show that if the scores for the true negatives and the scores for the true positives are considered as two samples to compare with the Mann-Whitney test, then the AUC can be computed from the corresponding Mann-Whitney statistic. Prove the formula that you have obtained.

[7] We assume that there are no ties, as the definition of the statistics is slightly more complicated if there are ties.