Semaine de pré-rentrée du master MVA
TD de Statistiques

Alexandre Bois

Septembre 2023

1 Multinomial random variables

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

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

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 $Z = (Z_1, \ldots, Z_K) \sim \mathcal{M}(\pi_1, \ldots, \pi_K, 1)$ then Z is a binary indicator vector with $\mathbb{P}(Z_k = 1) = \pi_k$.
2. If $Z^{(1)}, \ldots, Z^{(n)}$ is an i.i.d. sample from $\mathcal{M}(\pi_1, \ldots, \pi_K, 1)$ then defning $N_k = \sum_{i=1}^{n} Z_k^{(i)}$ k for all k, show that $N := (N_1, \ldots, N_K)$ follows the distribution $\mathcal{M}(\pi_1, \ldots, \pi_K, n)$.

2 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

$$DF(p, q) = F(p) - [F(q) + < F(q); p - q >]$$

1. Show that if F is the squared Euclidean norm in $\mathbb{R}^d$, the associated divergence is the squared Euclidean norm.
2. Consider two probability distributions $p = (p_i)_{1 \leq i \leq n}$ and $q = (q_i)_{1 \leq i \leq n}$ 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

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

   with the conventions $0/0 = 0$ and $0\mathsf{log}0 = 0$. Show that $KL(p, q)$ is the Bregman divergence $DH(p, q)$ associated with the entropy.
3. Show that $KL(p, q) \geq 0$ and $KL(p, q) = 0$ if and only if $p = q$.
   Remark : KL is not a distance as it is not symmetric and does not respect the triangle inequality.
4. Let $l : (X, \theta) \mapsto l(X, \theta)$ be a loss function and $R(\theta) = E[l(X, \theta)]$ the associated risk, where the expectation is taken w.r.t. the variable X. Denote by θ^* the minimizer of the risk, which is often called the target parameter, and consider the so-called excess risk $\mathcal{E}(\theta) := R(\theta) - R(\theta^*)$. Show that if the loss is strictly convex w.r.t. to its second argument and that R is differentiable, the excess risk can actually be interpreted as a Bregman divergence between θ and θ^*. What is the associated potential function ?

3 PCA

Let $x_1, \dots, x_n \in \mathbb{R}^p$. Let $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$. The empirical covariance matrix is

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^\top$$

We assume that the vectrors are centered, i.e. $\bar{x} = 0$.

1. Find the direction in $\mathbb{R}^p$ such that the variance along this direction is maximal, i.e. find :

$$v_1 = \underset{||v||_2=1}{\mathsf{argmax}}\ \mathsf{Var}((v^\top x_i)_{1\leq i\leq n})$$

2. Explain how to find the direction orthogonal to v_1 such that the variance along this direction is maximal.

Remark : for centered vectors, $\hat{\Sigma} = \frac{1}{n}X^\top X$, where X is the $n \times p$ matrix whose i^{th} row is x_i. So what we studied is actually the singular value decomposition (SVD) of $\frac{1}{\sqrt{n}}X$.

4 Method of moments vs maximum likelihood estimation

Recall : the beta distribution is a family of continuous probability distributions $(Beta(\alpha,\beta))_{\alpha,\beta>0}$ defined on the interval $[0,1]$, with densities $p_{\alpha,\beta}(x) \propto x^{\alpha-1}(1-x)^{\beta-1}$. Its mean and variance are respectively equal to $\frac{\alpha}{\alpha+\beta}$ and $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

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.

1. Compute the moment estimator $\hat{\theta}_{MO}$.
2. Show that the maximum likelihood estimator $\hat{\theta}_{MLE}$ exists and is unique and compute it.
3. Show that$\hat{\theta}_{MLE}$ follows a Beta distribution. What are the values of the parameters of this Beta ?
4. Deduce from the previous question the variance and the bias of the estimator.
5. What is the variance of the moment estimator ?
6. We consider the mean square error $E[(\theta - \hat{\theta})^2]$ as a measure of performance of the estimator. Compare the MSE for both estimators. Which estimator should be preferred ?

5 Maximum likelihood estimators

1. Compute the MLE $\hat{p}$ of p in the Bernoulli model : $X_i \sim Ber(p)$ iid. Then compute the asymptotic distribution of $\sqrt{n}(\hat{p} - p)$.
2. Compute the MLE $(\hat{m}, \hat{\sigma^2})$ of (m, σ^2) in the univariate Gaussian model : $X_i \sim \mathcal{N}(m, \sigma^2)$ iid. Then compute the asymptotic distribution of $\sqrt{n}(\hat{m} - m, \hat{\sigma^2} - \sigma^2)$.

6 Linear regression

The multiple linear regression model is :

$$Y = X\beta + \varepsilon$$

with $Y \in \mathbb{R}^n$, X is a $n \times p$ matrix, $\beta \in \mathbb{R}^p$ is the vector of unknown parameters, and $\varepsilon \in \mathbb{R}^n$. We assume that X is of rank p and ε is a centered random vector whose covariance matrix is $\sigma^2 I_n$. We study the least squares estimator $\hat{\beta}_{LSE} = \underset{\beta\in\mathbb{R}^p}{\mathsf{argmin}}||Y - X\beta||_2^2$.

1. Sow that $\hat{\beta}_{LSE} = (X^\top X)^{-1}X^\top Y$.
2. Compute the biais and variance of $\hat{\beta}_{LSE}$.

7 Bayesian estimation

Let $X = (X_1, \dots, X_n)$ be a sample of iid random variables such that $X_i \sim Ber(\theta)$ (Bernoulli distribution) with $\theta \in [0,1]$. Let $x = (x_1, \dots, x_n)$ be an observation of X. We consider the a priori distribution of θ to be the uniform distribution $\mathcal{U}([0,1])$. Compute the a posteriori distribution $p(\theta|x)$.

Liste de des références utilisées pour le cours.

Références

[1] A. Guyader. Statistique. *Sorbonne Univeristé (polycopié M1)*, 2023.

[2] R. Martin. Lecture notes on statistical theory. 2015.

[3] R. Martin. Lecture notes on advanced statistical theory. *Supplement to the lectures for Stat*, 511, 2016.

[4] L. Wasserman. *All of statistics : a concise course in statistical inference*, volume 26. Springer, 2004.