Lecture 7 : Statistics

Probability theory generally aims at predicting or explaining observations from a distribution model, on the contrary statistics aim at predicting the model or the parameters of the model from the observation. That is why, initially, Statistics were called “inverse probability”. There exists two school of statisticians: frequentist statistics and Bayesian statistics. Without going to much into details, one can merely say that in the frequentist approach, the parameters are deterministic values that we want to infer. On the contrary, under the Bayesian approach, the parameters are treated like any other variable and we try to express their probability law conditionally on the observation. Today the Bayesian approach seems to be predominant in the literature, its results are also arguably more precise and flexible, that is why we will mainly describe this approach in our course.

1 Bayesian statistics

1.1 Bayes Formula

Consider n independent copies $X_1, \dots, X_n$ of a continuous random vector $X : \Omega \to \mathcal{X}$, that we regroup in a data set $\mathcal{D} = \{X_1, \dots, X_n\}$. Given any other continuous random vector $\theta : \Omega \to \Theta$ (dependent or not on $\mathcal{D}$), one has the indentities:

$$p(\mathcal{D} = D, \theta = t) = p(\theta = t, \mathcal{D} = D) \quad \Longleftrightarrow \quad p(\mathcal{D} = D|\theta = t)p(\theta = t) = p(\theta = t|\mathcal{D} = D)p(\mathcal{D} = D), \tag{6.5}$$

that lead to the famous Bayes formula:

$$p(\theta = t|\mathcal{D} = D) = \frac{p(\mathcal{D} = D|\theta = t)p(\theta = t)}{p(\mathcal{D} = D)}.$$

Usually θ and $\mathcal{D}$ are ambiguously desgnating at the same time random vectors (as they were introduced) and elements of Θ, $\mathcal{X}^n$ then one typically work with the notations:

- $p(\theta) \equiv p_\theta(\theta)$
- $p(\mathcal{D}) \equiv p_\mathcal{D}(\mathcal{D})$
- $p(\theta|\mathcal{D}) \equiv p(\theta = \theta|\mathcal{D} = \mathcal{D})$
- $p(\mathcal{D}|\theta) \equiv p(\mathcal{D} = \mathcal{D}|\theta = \theta)$.

These notations are, arguably, highly ambiguous, however since their are widely use, and indeed quite practical, we will also adopt them for this course. The Bayes formula then writes:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int_\Theta p(\mathcal{D}|\theta)p(\theta)dt}, \tag{6.6}$$

To obtain the right-hand form, we integrated (6.5) on θ to get:

$$\int_\Theta p(\mathcal{D} = D|\theta = t)p(\theta = t)dt = \int_\Theta p(\theta = t|\mathcal{D} = D)p(\mathcal{D} = D)dt = p(\mathcal{D} = D),$$

thanks to Lemma 5.22. The form of the denominator in the right-hand term of (6.6) was to be expected since Lemma 5.22 imposes that $\int_\Theta p(\theta|\mathcal{D})d\theta = \int_\Theta p(\theta = t|\mathcal{D}) = \mathcal{D})dt = 1$.

Each of the densities appearing in (6.6) have classical denomination that we provide below.

- $p(\mathcal{D} = D)$: The evidence or marginal likelihood,
- $p(\theta)$: the prior
- $p(\mathcal{D} = D|\theta = t)$: the Likelihood
- $p(\theta|\mathcal{D})$: the Posterior.

Let us now describe this approach on a classical example.

Example 6.35 (Tossing coins)**.** *Given a coin, let us denote Y the random variable representing the outcome after tossing the coin ($Y \in \{0,1\}$) and $\theta \in [0,1]$ the probability of the coin coming up heads. If we toss the coin N times, recording outcomes in a training data set (it is a random vector) $\mathcal{D} = (Y_n : n = [N])$ whose entries are independent copies of Y, we seek $p(\theta|\mathcal{D})$.*

The likelihood expresses:

$$p(\mathcal{D}|\theta) = \prod_{n=1}^{N} \theta^{Y_n}(1-\theta)^{1-Y_n} = \theta^{N_1}(1-\theta)^{N_0},$$

where $N_1 = \sum_{n=1}^N \mathbb{I}(y_n = 1)$ and $N_0 = \sum_{n=1}^N \mathbb{I}(y_n = 0)$ (they are both discrete random variables) represent the number of heads and tails.

One can use an uninformative prior:

$$p(\theta) = Unif(\theta|0,1),$$

then the posterior would be proportional to the likelyhood.

More generally, bayesian statistician like to assume as prior:

$$p(\theta) = Beta(\theta|\breve{\alpha}, \breve{\beta}) \propto \theta^{\breve{\alpha}-1}(1-\theta)^{\breve{\beta}-1},$$

where $\breve{\alpha}$ and $\breve{\beta}$ are hyper-parameters. For $\breve{\alpha} = \breve{\beta} = 1$, we recover the uniform prior. We can think of the hyper-parameters as pseudocounts, analogous to empirical counts N_1 and N_0. The strength of the prior is controlled by $\breve{N} = \breve{\alpha} + \breve{\beta}$, known as the equivalent sample size. We can then compute the posterior by multiplying the likelihood by the prior:

$$p(\theta|D) \propto \theta^{N_1}(1-\theta)^{N_0}\theta^{\breve{\alpha}-1}(1-\theta)^{\breve{\beta}-1} \propto Beta(\theta|\breve{\alpha} + N_1, \breve{\beta} + N_0).$$

Since the posterior has the same form as the prior, it is called a conjugate prior.

1.2 Parameter's estimates

Definition 10 (MLE and MAP estimate)**.** *The Maximum likelihood estimate (MLE) is the parameter $\hat{\theta}_{MLE}$ defined as:*

$$\hat{\theta}_{MLE} = \arg\max_{\theta\in\Theta} p(\mathcal{D}|\theta)$$

The Maximum a posteriori (MAP) is the parameter $\hat{\theta}_{MAP}$ defined as:

$$\hat{\theta}_{MAP} = \arg\max_{\theta\in\Theta} p(\theta|\mathcal{D}) = \arg\max_{\theta} \log p(\theta) + \log p(D|\theta)$$

In other word, the MAP is the mode of the posterior distribution.

Note that the evidence $p(\mathcal{D})$ does not appear in the computations of $\hat{\theta}_{\mathrm{MLE}}$ and $\hat{\theta}_{\mathrm{MAP}}$ since it does not depends on θ.

Example 6.36. *Let us return to the setting of Exercise 6.35. Calculus yields:*

$$\hat{\theta}_{MAP} = \frac{\breve{\alpha} + N_1 - 1}{\breve{\alpha} + N_1 - 1 + \breve{\beta} + N_0 - 1}.$$

If we use a uniform prior, $p(\theta) \propto 1$, the MAP estimate becomes the MLE, since $\log p(\theta) = 0$:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \log p(D|\theta) = \frac{N_1}{N_1 + N_0} = \frac{N_1}{N}$$

This is intuitive and easy to compute. However, the MLE can be very misleading in the small sample setting. For example, suppose we toss the coins N times, but never see any heads, so $N_1 = 0$. In this case, we would estimate that $\hat{\theta} = 0$, which means we would not predict any future observations to be heads either. This is a very extreme estimate, that is likely due to insufficient data. We can solve this problem using a MAP estimate with a stronger prior. For example, if we use a Beta$(\theta|2, 2)$ prior, we get the estimate

$$\hat{\theta}_{MAP} = \frac{N_1 + 1}{N_1 + 1 + N_0 + 1} = \frac{N_1 + 1}{N + 2}$$

This is called add-one smoothing.

The posterior mode can be a poor summary of the posterior, since it corresponds to picking a single point from the entire distribution. The posterior mean is a more robust estimate, since it is a summary statistic derived by integrating over the distribution, $\bar{\theta} = \int \theta p(\theta|D) d\theta$. In the case of a beta posterior, $p(\theta|D) = Beta(\theta|\alpha, \beta)$, the posterior mean is given by

$$\bar{\theta} \triangleq \mathbb{E}[\theta|D] = \frac{\hat{\alpha}}{\hat{\beta} + \hat{\alpha}} = \frac{\hat{\alpha}}{\hat{N}}$$

where $\hat{N} = \hat{\beta} + \hat{\alpha}$ is the strength (equivalent sample size) of the posterior.

We will now show that the posterior mean is a convex combination of the prior mean, $m = \frac{\alpha}{\breve{N}}$ and the MLE, $\hat{\theta}_{MLE} = \frac{N_1}{N}$:

$$\mathbb{E}[\theta|D] = \frac{\breve{\alpha} + N_1}{\breve{\alpha} + N_1 + \breve{\beta} + N_0} = \frac{\breve{\alpha}}{\breve{N} + N} + \frac{N_1}{\breve{N} + N} = \frac{\breve{N}}{\breve{N} + N} m + \frac{N}{\breve{N} + N} \hat{\theta}_{MLE} = \lambda m + (1 - \lambda) \hat{\theta}_{MLE}$$

where $\lambda = \frac{\hat{N}}{\hat{N} + \breve{N}}$ is the ratio of the prior to posterior equivalent sample size. We see that the weaker the prior, the smaller λ, and hence the closer the posterior mean is to the MLE.

1.3 Credible intervals

A posterior distribution is (usually) a high dimensional object that is hard to visualize and work with. A common way to summarize such a distribution is to compute a point estimate, such as the posterior mean or mode, and then to compute a *credible interval*, which quantifies the uncertainty associated with that estimate. (A credible interval is not the same as a confidence interval, which is a concept from frequentist statistics which we discuss in Section 3.3.5.1.)

More precisely, we define a $100(1-\alpha)\%$ credible interval to be a (contiguous) region $C = (\ell, u)$ (standing for lower and upper) which contains $1 - \alpha$ of the posterior probability mass, i.e.,

$$C_{\alpha}(\mathcal{D}) = [\ell, u], \qquad \text{where: } P(\ell \leq \theta \leq u|\mathcal{D}) = 1 - \alpha \tag{6.7}$$

There may be many intervals that satisfy(6.7), so we usually choose one such that there is $(1-\alpha)/2$ mass in each tail; this is called a *central interval*. If the posterior has a known functional form, we can compute the posterior central interval using $\ell = F^{-1}(\alpha/2)$ and $u = F^{-1}(1-\alpha/2)$, where F is the cdf of the posterior, and F^{-1} is the inverse cdf. For example, if the posterior is Gaussian, $p(\theta|\mathcal{D}) = \mathcal{N}(0,1)$, and $\alpha = 0.05$, then we have $\ell = \Phi^{-1}(\alpha/2) = -1.96$, and $u = \Phi^{-1}(1-\alpha/2) = 1.96$, where Φ denotes the cdf of the Gaussian. This justifies the common practice of quoting a credible interval in the form of $\mu \pm 2\sigma$, where μ represents the posterior mean, σ represents the posterior standard deviation, and 2 is a good approximation to 1.96.

A problem with central intervals is that there might be points outside the central interval which have higher probability than points that are inside. This motivates an alternative quantity known as the *highest posterior density* or *HPD* region, which is the set of points which have a probability above some threshold. More precisely we find the threshold p^* on the pdf such that

$$1-\alpha = \int_{\theta : p(\theta|\mathcal{D}) \geq p^*} p(\theta|\mathcal{D}) d\theta$$

and then define the HPD as

$$C_\alpha(\mathcal{D}) = \{\theta : p(\theta|\mathcal{D}) \geq p^*\}$$

2 The problem comes first

In this section, the attention is not put on the model and its parameters that might be to complicated too be assessed but rather on a general classification of regression task. We present in the table below the main objects of our framework:

Dataset

$\mathcal{D} = (X_i, Y_i)_{i=1,\cdots,n}$: Training set with n independent samples, all having the same distribution as (X, Y).

X : Predictors, features or inputs of a sample. $X \in \mathcal{X} \subset \mathbb{R}^p$ is a p-length (random) vector.

Y : Response or outcome of a sample. $Y \in \mathcal{Y} \subset \mathbb{R}^k$ is a k-length vector.

Learning paradigm

$f(\cdot)$: A decision function. $f : \mathcal{X} \to \mathbb{R}^k; x \to f(x)$ maps inputs (feature) space to the outcome-space, say the decision function is $f(x)$ given a sample $X = x$.

$l(\cdot,\cdot)$: A loss function. $l : \mathcal{Y} \times \mathbb{R}^k \to \mathbb{R}; \big(y, f(x)\big) \to l\big(y, f(x)\big)$ measure the discrepancy between the true outcome and the decision function.

$R(f)$: The risk of the decision function.

$$R(f) \equiv \mathbb{E}\Big(l\big(Y, f(X)\big)\Big)$$

Note that in the definition of the **risk** function (or **generalization error**) $(R(f) = \mathbb{E}\left(l\left(Y, f(X)\right)\right))$ the expectation is taken w.r.t. both X and Y. Given a testing dataset $\mathcal{T}_m = (X_j^{\text{Tst}}, Y_j^{\text{Tst}})_{j=1,\cdots,m}$, the risk function is empirically evaluated as an averaged loss:

$$\widehat{R}_m(f) = \frac{1}{m}\sum_{j=1}^{m} l\left(y_j^{\text{Tst}}, f(x_j^{\text{Tst}})\right).$$

The risk function can be used to check the performance of a decision function, yet we want to further investigate its "efficiency". To this end, we first introduce the best decision function, namely Bayes decision function (rule), then compute the discrepancy to measure "efficiency".

Definition 11 (Bayes decision rule)**.** *A Bayes (decision) rule is defined as the smallest risk achievable by any measurable decision function, that is,*

$$f^* = \arg\min R(f),$$

where the minimum is taken over all possible measurable functions.

We illustrate the risk function and its Bayes rule with two loss example, namely, the mis-classification error (MCE) and the mean square error (MSE).

Definition 12 (MSE)**.** *The mis-classification error (MCE) in binary classification (*$Y \in \{-1, +1\}$*) is defined as:*

$$R(f) = \mathbb{P}\left(Y \neq \mathit{Sgn}(f(X))\right) = \mathbb{E}\left[\mathbb{1}\left(Y \neq \mathit{Sgn}(f(X))\right)\right] = \mathbb{E}\left[\mathbb{1}\left(Yf(X) \leq 0\right)\right],$$

Lemma 6.37 (Mis-classification error)**.** *In a binary classification problem, f^* is a Bayes rule iff*

$$\mathit{Sgn}\left(f^*(x)\right) = \mathit{Sgn}\left(\mathbb{P}(Y = 1|X = x) - 1/2\right).$$

Proof. For any binary classifier f and any input x, let us first compute the probability of misclassifying x. Since $f(X)$ and Y are independent conditionally on X, one can compute:

$$\begin{aligned}
\mathbb{P}(f(X) \neq Y|X = x) &= \mathbb{P}(f(X) = 1, Y = -1|X = x) + \mathbb{P}(f(X) = -1, Y = 1|X = x) \\
&= \mathbb{P}(f(X) = 1|X = x)\mathbb{P}(Y = -1|X = x) + \mathbb{P}(f(X) = -1|X = x)\mathbb{P}(Y = 1|X = x) \\
&= \mathbb{1}_{f(x)=1}\mathbb{P}(Y = -1|X = x) + \mathbb{1}_{f(x)=-1}\mathbb{P}(Y = 1|X = x) \\
&= \mathbb{1}_{f(x)=1}[1 - \mathbb{P}(Y = 1|X = x)] + [1 - \mathbb{1}_{f(x)=1}]\mathbb{P}(Y = 1|X = x) \\
&= [1 - 2\mathbb{P}(Y = 1|X = x)]\mathbb{1}_{f(x)=1} + \mathbb{P}(Y = 1|X = x).
\end{aligned}$$

Therefore:

$$\begin{aligned}
P(f(X) \neq Y|X = x) - P(t(X) \neq Y|X = x) &= [1 - 2P(Y = 1|X = x)]\mathbb{1}_{f(x)=1} + P(Y = 1|X = x) \\
&\quad - [1 - 2P(Y = 1|X = x)]\mathbb{1}_{t(x)=1} - P(Y = 1|X = x) \\
&= [1 - 2P(Y = 1|X = x)][\mathbb{1}_{f(x)=1} - \mathbb{1}_{t(x)=1}].
\end{aligned}$$

We wrote this quantity as a product of two terms, so that its sign depends on the signs of the two terms.

- If $1 - 2P(Y = 1|X = x) > 0$, then $P(Y = 1|X = x) < \frac{1}{2}$ and, $t(x) = -1$. Thus, $\mathbb{1}_{t(x)=1} = 0$ and in this case the second term, $\mathbb{1}_{f(x)=1} - \mathbb{1}_{t(x)=1} = \mathbb{1}_{f(x)=1} \in \{0, 1\}$, is also positive.
- If $1 - 2P(Y = 1|X = x) < 0$, then $P(Y = 1|X = x) > \frac{1}{2}$ and, $t(x) = +1$. Thus, $\mathbb{1}_{t(x)=1} = 1$ and in this case the second term, $\mathbb{1}_{f(x)=1} - \mathbb{1}_{t(x)=1} = \mathbb{1}_{f(x)=1} - 1 \in \{-1, 0\}$, is also negative.

Therefore, the quantity above is positive, since it is the product of either two positive numbers or two negative numbers. That finally implies that for all $f : \mathcal{X} \to \{0, 1\}$, measurable:

$$P(f(X) \neq Y|X = x) \geq P(t(X) \neq Y|X = x)$$

□

In a regression task, one usually rather works with the so called mean square error.

Definition 13 (MSE)**.** *The mean squared error (MSE) in (multi-outcome) regression (*$Y \in \mathbb{R}^k$*) is defined as:*

$$R(f) = \mathbb{E}\left[(Y - f(X))^2\right],$$

Lemma 6.38. *The Bayes rule for the mean-square error is defined as:*

$$f^*(x) = \mathbb{E}\left[Y|X = x\right].$$

Proof. Given a decision function $f : \mathcal{X} \to \mathbb{R}^k$, the risk for the MSE writes:

$$\begin{aligned} R(f) &= \mathbb{E}\left[(Y - f(X))^2\right] = \mathbb{E}\left[\mathbb{E}\left[(Y - f(X))^2 \ \mid X\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[(Y - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - f(X))^2 \ \mid X\right]\right] \end{aligned}$$

Now, we know from Lemma 5.31 that $Y - \mathbb{E}[Y|X]$ is independent with $\mathbb{E}[Y|X] - f(X)$ conditionally on X, then a similar result as Lemma 5.34 (just use the fact that $\mathbb{E}[\mathbb{E}[Y|X] - f(X) \mid X] = 0$) allows us to set:

$$\begin{aligned} R(f) &= \mathbb{E}\left[\mathbb{E}\left[(Y - \mathbb{E}[Y|X])^2 \ \mid X\right] + \mathbb{E}\left[(\mathbb{E}[Y|X] - f(X))^2 \ \mid X\right]\right] \\ &= \mathbb{E}\left[(Y - f^*(X))^2\right] + \mathbb{E}\left[(f^*(X) - f(X))^2\right]. \end{aligned}$$

One can then conclude that $R(f) \geq R(f^*)$ since $\mathbb{E}\left[(f^*(X) - f(X))^2\right] \geq 0$. □