Statistics
MDI220
Introduction

Thomas Bonald
Institut Polytechnique de Paris

2023 – 2024

Statistics is the science of data. The objective is to extract useful information from observations so as to make predictions. The approach consists in viewing each data sample x as the value of a random variable X.

1 Parametric model

A statistical model is *parametric* if the probability distribution of X belongs to some family of distributions indexed by some parameter θ of finite dimension.

A parametric model is a set $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ of probability distributions with $\Theta \subset \mathbb{R}^k$ for some $k \geq 1$.

The parameter θ is unknown and must be learned, using the observation x. This is possible only if each probability distribution $P_\theta \in \mathcal{P}$ is defined by a unique parameter θ.

The parametric model $\mathcal{P}$ is said to be *identifiable* if the mapping $\theta \mapsto P_\theta$ is a bijection.

In most practical cases, the probability distribution P_θ has a density with respect to some measure ν (e.g., the Lebesgue measure).

The parametric model $\mathcal{P}$ is said to be *dominated* if there exists some measure ν such that P_θ has a density p_θ with respect to ν for all $\theta \in \Theta$

Example 1 *The model $\mathcal{P} = \{P_\theta \sim \mathcal{N}(\mu, \sigma^2), \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+\}$ of Gaussian distributions is dominated by the Lebesgue measure $\nu = \lambda$. The density is $p_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.*

Example 2 *The model $\mathcal{P} = \{P_\theta \sim \mathcal{B}(\theta), \theta \in [0,1]\}$ of Bernoulli distributions is dominated by the counting measure $\nu = \delta_0 + \delta_1$. The density is $p_\theta(x) = \theta^x (1-\theta)^{1-x}$.*

When n observations are available, say $x_1, \ldots, x_n$, it is common to assume that these observations are independent and identically distributed. If the model is dominated, the probability density function of the vector of observations $x = (x_1, \ldots, x_n)$ is the product of the probability density functions. We use the same notation p_θ for convenience.

For a dominated model with n independent observations $x = (x_1, \dots, x_n)$, $p_\theta(x) = p_\theta(x_1) \dots p_\theta(x_n)$.

2 Decision function

The role of the statistician is to take *actions* based on the observations.

A decision function δ is a mapping from the set of observations $\mathcal{X}$ to the set of actions $\mathcal{A}$.

The set of actions $\mathcal{A}$ depends on the task:

Estimation → The objective is to find θ and thus $\mathcal{A} = \Theta$.

Hypothesis testing → The objective is to answer a binary question and thus $\mathcal{A} = \{0, 1\}$.

Confidence region → The objective is to find a confidence region for θ and thus $\mathcal{A} = \mathcal{P}(\Theta)$.

3 Loss function

A loss function is used to assess the quality of an action $a \in \mathcal{A}$. The quality of this action depends on the parameter θ.

A loss function L is a mapping from $\Theta \times \mathcal{A}$ to $\mathbb{R}_+$ so that $L(\theta, a)$ is the cost of selecting action a for the parameter θ.

Usual loss functions are:

Estimation → For $\theta, a \in \mathbb{R}$, the square error $L(\theta, a) = (\theta - a)^2$.

Confidence region → For $\theta \in \mathbb{R}$, $a \subset \mathbb{R}$, the fit error $L(\theta, a) = 1_{\{\theta \notin a\}}$.

4 Risk

To assess the quality of a decision function δ, it is necessary to consider all possible actions taken by this decision function, depending on the observations. This can be done by averaging the loss function over all observations (and thus all actions taken for these observations).

The risk R is the expectation of the loss function L over all observations, $R(\theta, \delta) = \mathrm{E}_\theta(L(\theta, \delta(X)))$.

Note that the risk R depends on the parameter θ, which is unknown. Several strategies exist to select the best decision function:

Min-max → Minimize the maximum risk, $\max_{\theta \in \Theta} R(\theta, \delta)$.

Bayesian approach → Minimize the expected risk $\mathrm{E}(R(\theta, \delta))$, assuming some prior distribution on θ.