

Review Lab 2

Machine Learning I - UMONS

February 2022

- When we conduct an experiment, we look at the outcome of a stochastic process taking values in some sample space Ω .
 - Defining all possible outcomes of the experiment.
- An event \mathbf{A} is a subset of Ω ($\mathbf{A} \in \Omega$). An event \mathbf{A} occurs if the outcome of the experiment belongs to \mathbf{A} .
- A **random variable** X is a mapping from the sample space Ω to the reals.
 - E.g. $X = \# \text{heads from throwing a coin 10 times.}$
 - $X \in \{0, 1, 2, \dots, 10\}$

- Two kinds of random variables :

- **Discrete** random variables

- Support of X is discrete : $\mathcal{X} \in \{0, 1, 2, 3, \dots\}$
 - Associated to a probability mass function (pmf) $p_X(x)$:

$$p_X(x) = \mathbb{P}(X = x)$$

- $p_X(x) \geq 0, \forall x \in X$
 - $\sum_{x \in X} p_X(x) = 1$

- **Continuous** random variables :

- Support of X is continuous.
 - Associated to a probability density function $f_X(x)$:

$$\int_a^b f_X(x) dx = \mathbb{P}(a \leq x \leq b)$$

- $f_X(x) \geq 0, \forall x \in X$
 - $\int_X f_X(x) dx = 1$

- **Expectation** of a discrete random variable :

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x p_X(x) = \mu_X$$

- **Expectation** of a continuous random variable :

$$\mathbb{E}[X] = \int_{\mathcal{X}} x f_X(x) dx = \mu_X$$

- Properties of the expectation :

- For any constant c , $\mathbb{E}[X + c] = \mathbb{E}[X] + c$
- For any constant c , $\mathbb{E}[cX] = c\mathbb{E}[X]$
- For any function g :
 - $\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) p_X(x)$ for discrete variables.
 - $\mathbb{E}[g(X)] = \int_{\mathcal{X}} g(x) f_X(x) dx$ for continuous variables.
- For any functions g and h , $\mathbb{E}[g(X) + h(X)] = \mathbb{E}[g(X)] + \mathbb{E}[h(X)]$

- **Variance** of a random variable :

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[(X - \mu_X)^2]\end{aligned}$$

- **Standard deviation** of a random variable :

$$\sigma(X) = \sqrt{\text{Var}(X)}$$

- Properties of the variance :

- $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
- For any constant c , $\text{Var}(cX) = c^2\text{Var}(X)$
- For any constant c , $\text{Var}(c + X) = \text{Var}(X)$

- Given two discrete random variables X and Y , their **joint** pmf is written:

$$p_{XY}(x, y) = \mathbb{P}(X = x, Y = y)$$

- Given two continuous random variables X and Y , their **joint** pdf is written $f_{XY}(x, y)$ such that:

$$\int_a^b \int_c^d f_{XY}(x, y) dx dy = \mathbb{P}(a \leq x \leq b, c \leq y \leq d)$$

- The **marginal** pmf of X is defined as :

$$p_X(x) = \sum_{y \in \mathcal{Y}} p_{XY}(x, y)$$

- The **marginal** pdf of X is defined as :

$$f_X(x) = \int_{\mathcal{Y}} f_{XY}(x, y) dy$$

- For any function g , the joint expectation is defined as :

- For discrete random variables :

$$\mathbb{E}[g(X, Y)] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} g(x, y) p_{XY}(x, y)$$

- For continuous random variables :

$$\mathbb{E}[g(X, Y)] = \int_{\mathcal{X}} \int_{\mathcal{Y}} g(x, y) f_{XY}(x, y) dx dy$$

- The covariance of two random variables X and Y is defined as :

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \end{aligned}$$

- Useful properties :

- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
 - $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$

- The **conditional** pmf of Y given X is :

$$p_{Y|X}(y|x) = \frac{p_{XY}(x, y)}{p_X(x)}$$

- The **conditional** pdf of Y given X is :

$$f_{Y|X}(x|y) = \frac{f_{XY}(x, y)}{f_X(x)}$$

- The **law of total probability** for discrete random variables gives :

$$\begin{aligned} p_X(x) &= \sum_{y \in \mathcal{Y}} p_{XY}(x, y) \\ &= \sum_{y \in \mathcal{Y}} p_{X|Y}(x|y) p_Y(y) \end{aligned}$$

- **Bayes' rule** :

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x) p_X(x)}{\sum_{x \in \mathcal{X}} p_{Y|X}(y|x) p_X(x)}$$

- Replace pmf's by pdf's and sums by integrals for continuous random variables.

- The **conditional expectation** of Y given X for discrete random variables is:

$$\mathbb{E}[Y|X = x] = \sum_{y \in \mathcal{Y}} yp_{Y|X}(y|x)$$

- The **conditional expectation** of Y given X for continuous random variables is :

$$\mathbb{E}[Y|X = x] = \int_{\mathcal{Y}} yf_{Y|X}(y|x)dy$$

- The law of **total expectation** yields :

$$\mathbb{E}[Y] = \sum_{x \in X} \mathbb{E}[Y|X = x]p_X(x) \quad \text{or} \quad \mathbb{E}[Y] = \int_X \mathbb{E}[Y|X = x]f_X(x)dx$$

- Two random variables X and Y are independent i.i.f :

$$p_{XY}(x, y) = p_X(x)p_Y(y) \quad \text{or} \quad f_{XY}(x, y) = f_X(x)f_Y(y)$$

- If two random variables X and Y are independent, then :

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

- Usually, we don't have access to the entire population of a random variable X .
 - The population statistics, such as the mean μ_X and the variance $\text{Var}(X)$ of p_X are unknown !
 - We must rely on **point estimators** for these quantities given a finite number of samples $X_1, \dots, X_n \sim p_X$.
 - Ex : The sample mean, $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, is an estimator of μ_X .
- A central concept in statistical modeling consists in supposing that some observed data y_1, \dots, y_n originated from a distribution p_Y .
 - We don't know p_Y , but we want to estimate it.
 - We suppose that the data originated from a distribution $p(y; \theta)$, and we want to find the best θ such that $p(y; \theta)$ is as close as possible to p_Y .

- We want to maximize the **likelihood** that $p(y; \theta)$ generated the observed samples y_1, \dots, y_n .
- We make the hypothesis that the variables are **independent and identically distributed (i.i.d)**. The likelihood function is defined as :

$$\begin{aligned} L(\theta) &= p(y_1, \dots, y_n; \theta) \\ &= \prod_{i=1}^n p(y_i; \theta) \end{aligned}$$

- We want to find the **Maximum Likelihood Estimator (MLE)**, i.e. the value of θ that maximizes the likelihood function :

$$\begin{aligned} MLE = \hat{\theta} &= \operatorname{argmax}_{\theta \in \Theta} L(\theta) \\ &= \operatorname{argmax}_{\theta \in \Theta} \log L(\theta) \\ &= \operatorname{argmax}_{\theta \in \Theta} \sum_{i=1}^n \log p(y_i; \theta) \end{aligned}$$

- Taking the first derivative of $\log L(\theta)$ with respect to θ , equalling it to zero and solving for θ yields the MLE :

$$(\log L)'(\theta) = 0$$

- We can further check that this is indeed a maximum by taking the second derivative of $\log L(\theta)$ with respect to θ and verifying that :

$$(\log L)''(\theta) \leq 0$$