LEPL1109 Statistics and data science

Simon Desmidt

Academic year 2023-2024 - Q1

UCLouvain

Contents

Reminders

1.1 Random variable

A random variable is a measurable function from the sample space Ω to the set of real numbers $\mathbb{R}$, $X : \Omega \to \mathbb{R}$. We denote it by a capital letter and its realization by a lowercase letter : $X(\omega) = x$.

- The state space of a rv X is the set of all possible values of the rv : $\{x \in \mathbb{R} | x = X(\omega), \omega \in \Omega\}$.
- A rv is called discrete if its state space has a finite or countable number of elements.
- A rv is called continuous if it takes arbitrary real values between a minimum and a maximum. The state is either an interval of $\mathbb{R}$, or $\mathbb{R}$.

1.2 Probability distribution

The probability mass function (pmf) $p(x)$ of a discrete random variable X is a function that associates to all values x of X a probability : $p(x) = P(X = x)$. If x is in the range of X, then $p(x) > 0$ and $p(x) = 0$ if not. Furthermore, $\sum_i p(x_i) = 1$.

1.3 Probability density function

Let X be a continuous random variable. The probability density function (pdf) of X is the function $f(x) \geq 0$ such that for any $I \subset \mathbb{R}$, we have $P(X \in I) = \int_I f(x)dx$.

1.4 Cumulative distribution function

The cumulative distribution function $F(x)$ of a random variable X indicate for each possible value of x the probability that X takes the value equal or less than x : $F(x) = P(X \leq x)$. The pdf is therefore the x-derivative of the cdf.

1.5 Expectation

The expectation of a rv X is noted $\mathbb{E}(X) = \mu_X$. If X is a discrete rv, then

$$\mu_X = \sum_{x \in range(X)} xp(x) \tag{1.1}$$

If X is continuous with a density $f(x)$, the expectation of X is

$$\mu_X = \int_{-\infty}^{\infty} x f(x) dx \tag{1.2}$$

1.5.1 Properties

- $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
- $\mathbb{E}(aX+b) = a\mathbb{E}(X) + b$[1]

1.5.2 Composition

The expected value of a real-valued function h of a discrete rv X is

$$\mathbb{E}(h(X)) = \sum_{x \in range(X)} h(x)p(x) \tag{1.3}$$

If X is continuous with a density $f(x)$, the expectation of $h(X)$ is

$$\mathbb{E}(h(X)) = \int_{-\infty}^{\infty} h(x) f(x) dx \tag{1.4}$$

- $\mathbb{E}(h(X)) \neq h(\mathbb{E}(x))$
- $\mathbb{E}(XY) \neq \mathbb{E}(X)\mathbb{E}(Y)$

1.6 Variance

The variance of a random variable X is denoted by $\mathbb{V}(X) = \sigma_X^2$ and is defined as $\mathbb{V}(X) = \mathbb{E}\left((X - \mu_X)^2\right)$. The standard deviation is $\sigma_X = \sqrt{\mathbb{V}(X)}$.

1.6.1 Properties

If X and Y are two rv and $a, b \in \mathbb{R}$,

- $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X)$
- $\mathbb{V}(a) = 0$
- $\mathbb{V}(a + bX) = b^2\mathbb{V}(X)$
- $\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y)$ is true iif X and Y are independent.

$\rightarrow$ N.B. : $\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X dx$

1.7 Law of Large Numbers

Let $X_{i=1,\ldots,n}$ be a sequence of uncorrelated rv's with the same expectation $\mu_X = \mathbb{E}(X_i)$ and variance σ_X^2. When $n \rightarrow \infty$, the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges in probability to μ_X.

[1] This is due to the properties of the integral.

1.8 Quantiles of a distribution

Let p be a probability between 0 and 1 and X be a rv. The number q_p that satisfies the relation $P(X \leq q_p) = p$ is the quantile of ordre p for X. If X is continuous and if its cdf $F(x)$ is invertible, then $q_p = F^{-1}(p)$.|

1.9 Function of random variables

Let X be a continuous rv with density $f_X(x)$. The pdf of $Y = a + bX$ is given by

$$f_Y(y) = \frac{1}{|b|} f_X\left(\frac{y-a}{b}\right) \tag{1.5}$$

→ N.B. : The Gamma function is defined by $\Gamma : \mathbb{Z} \to \mathbb{Z} : z \to (z-1)!$.

Independence and linear dependence

2.1 Independant random variables

Two rv's X and Y are independent $X \perp\!\!\!\perp Y$ if for every $A, B \in \mathbb{R}$, we have $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$.
The consequences are the following :

- $p(x, y) = p_X(x)p_Y(y)$ when the rv's are discrete
- $f(x, y) = f_X(x)f_Y(y)$ when the rv's are continuous
- $\mathbb{C}(X, Y) = \mathbb{C}(Y, X)$
- $\mathbb{C}(X, X) = \mathbb{V}(X)$ and $\mathbb{C}(a, X) = 0 \; \forall a \in \mathbb{R}$
- if $X \perp\!\!\!\perp Y$, then $C(X, Y) = 0$
- $\mathbb{C}(aX + bY, Z) = a\mathbb{C}(X, Z) + b\mathbb{C}(Y, Z)$
- $\mathbb{V}(aX + bY) = a^2\mathbb{V}(X) + b^2\mathbb{V}(Y) + 2ab\mathbb{C}(X, Y)$

2.2 Covariance

Let X, Y be rv's. The covariance between them is

$$\sigma_{XY} = Cov(X, Y) = \mathbb{E}\left((X - \mu_X)(Y - \mu_Y)\right) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \tag{2.1}$$

If we observe n outcomes $(x_k)_{k=1,\ldots,n}$ and $(y_k)_{k=1,\ldots,n}$ of rv's X, Y, the empirical covariance is estimated as follows :

$$\sigma_{XY} \approx s_{XY} = \frac{1}{n-1}\sum_{n=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \tag{2.2}$$

If $\mathbb{C} > 0$, then the rv's move in the same direction, and they move in opposite directions if $\mathbb{C} < 0$.

2.3 Correlation

The correlation is a measure of the linear dependence between two rv's easier to interpret than σ_{XY} because it has no unit and is in $[-1, 1]$.

Let X, Y be rv's. The correlation between them is defined as

$$\rho_{XY} = \frac{\mathbb{C}(X, Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{2.3}$$

- ρ_{XY} is scale invariant.
- $-1 \leq \rho XY \leq 1$
- If $\rho_{XY} = 1$, then $Y = a + bX$ with $b \in \mathbb{R}^+$.
- If $\rho_{XY} = -1$, then $Y = a + bX$ with $b \in \mathbb{R}^-$.

If we observe n outcomes $x_1, ..., x_n$ and $y_1, ..., y_n$ of the rv's, the correlation is estimated by the empirical correlation :

$$\rho_{XY} \approx r_{XY} = \frac{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{2.4}$$

Normal random variable and Central Limit Theorem

3.1 Normal distribution

A rv X follows a normal distribution if its density is given by the following function :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{3.1}$$

where $\mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$ are parameters[1]. We note $X \sim N(\mu, \sigma^2)$.

The moment generating function is defined by the relation $\frac{\partial^n m_X(t)}{\partial t^n}|_{t=0} = \mathbb{E}(X^n)$. Therefore, for a normal distribution, it is given by

$$m_X(t) \coloneqq \mathbb{E}(e^{tX}) = \exp\left(\mu t + \frac{1}{2}t^2\sigma^2\right) \tag{3.2}$$

3.1.1 Standardization

- If $X \sim N(\mu, \sigma^2)$, then for any $a, b \neq 0$, $a + bX$ is a normal rv $N(a + b\mu, b^2\sigma^2)$.

- If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ and the covariance is $\sigma_{XY} = \mathbb{C}(X, Y)$, then

$$aX + bY \sim N(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}) \tag{3.3}$$

- If $X \sim N(\mu_X, \sigma_X^2)$, then $\frac{X-\mu}{\sigma} \sim Z = N(0, 1)$ and

$$P(X \le x) = P\left(Z \le \frac{x-\mu}{\sigma}\right) \tag{3.4}$$

3.2 Central Limit Theorem

Let $X_1, ..., X_n$ be a sequence of independent identically distributed (iid) rv's (not following any particular distribution) with $\mathbb{E}(X_i) = \mu$ and $\mathbb{V}(X_i) = \sigma^2$. As $n \to \infty$,

$$Z_n = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \to Z \sim N(0, 1) \iff P(Z_n \le z) \to P(Z \le z)\ \forall z \tag{3.5}$$

[1] $\mu = \mathbb{E}(X)$ and $\sigma^2 = \mathbb{V}(X)$

This means that, for a large n, the distribution of the mean $\bar{X}_n$ and the sum $S_n = \sum_{i=1}^n X_i$ may be approached by a normal distribution

$$\begin{cases} \bar{X}_n \sim N(\mu, \sigma^2/n) \\ S_n \sim N(n\mu, n\sigma^2) \end{cases} \tag{3.6}$$

3.3 Chi-square distribution

The chi-square (χ^2) distribution with n degrees of freedom is the distribution of a sum of squares of n independent standard normal $N(0,1)$ rv's.

A rv X defined on $\mathbb{R}^+$ follows a χ^2-distribution of parameter n when its density is given by

$$\frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} \exp -x/2 \tag{3.7}$$

Its expectation is n and its variance is $2n$.

3.4 Student's T

Let Z and Y be two independent rv $Z \sim N(0,1)$ and $Y \sim \chi^2_n$. Then we define $T_n = \frac{Z}{\sqrt{Y/n}}$ as the Student's T rv with n degrees of freedom. Its density is

$$f_{T_n}(t) = \frac{\Gamma(\frac{n+1}{2}}{\sqrt{n\pi}\Gamma(n/2)} \left(1 + \frac{t^2}{2}\right)^{-\frac{n+1}{2}} \quad \text{with } t \in \mathbb{R} \tag{3.8}$$

Estimation

The collection of rv's X_i is called a random sample of size n if they have the same probability distribution $f(x|\theta)^1$ and are mutually independent.

4.1 Estimator

If we assume that $X_i \sim N(\mu, \sigma)$, then $\sigma = (\mu, \sigma) \in \Theta = \mathbb{R}_+^2$ and we have to estimate the parameters μ, σ.

An estimator of θ, generically denoted by $\hat{\theta}$, is any function $h(\cdot)$ of the random sample $\hat{\theta} = h(X_1, ..., X_n) \in \Theta$ used to estimate θ.

An estimate of θ is an observed value of this estimator calculated from the observed sample $x_1, ...x_{,n} : \hat{\theta}_{obs} = h(x_1, ..., x_n) \in \Theta$. µ

→ N.B. : the estimator is a function of n random variables and therefore is also a random variable.

4.2 Bias

$\hat{\theta}$ is an unbiased estimator of θ if

$$\mathbb{E}(\hat{\theta}) = \mathbb{E}(h(X_1, ..., X_n)) = \theta \tag{4.1}$$

The bias is the difference between the expectation and the real unknown value : $B(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta)$.

The mean square error (MSE) measures the average error : $MSE(\hat{\theta}) = \mathbb{E}\left(\left(\hat{\theta} - \theta\right)^2\right)$.

We can also define the bias-variance with the following relation :

$$MSE(\hat{\theta}) = B^2(\hat{\theta}) + \mathbb{V}(\hat{\theta}) \tag{4.2}$$

[1] θ being a vector of parameters : $\theta \in \Theta$

4.3 Method of moments

We observe $x_1, ..., x_n$ realizations of $X_{1:n}$. We think that $X_{1:n}$ have the same pdf $f(x|\theta)$ as X. In order to estimate $\theta \in \mathbb{R}^d$, we match the d moments with the d empirical moments

$$\mu_k(\theta) = M_k \iff \mathbb{E}\left(X^k\right) = \frac{1}{n}\sum_{i=1}^{n} X_i^k \; k = 1, ..., d \tag{4.3}$$

The value of the estimator converges to the true value when $n \to \infty$. Other than that, we do not know much about it, but we can say that the moment estimators are easy to construct, though do not always possess the best statistical properties.

4.4 Likelihood maximization

Let us consider a random sample $X_{1:n} \sim X$. We think that X has a pdf $f(x|\theta)$. We know that

$$P(x \le X \le x + dx) \approx f(x|\theta)dx \tag{4.4}$$

Since $X_{1:n}$ are independent, the probability to observe realizations $x_{1:n}$ is

$$P(x_i \le X_i \le x_i + dx \; \forall i) = \prod_{k=1}^{n} f(x_k|\theta)dx \tag{4.5}$$

The probability that the observed sample has been generated by the model is then proportional to the likelihood function

$$L(x_i|\theta) := \prod_{k=1}^{n} f(x_k|\theta) \tag{4.6}$$

The maximum likelihood estimator (MLE) of θ is the value which maximises the likelihood of the observed sample : $\hat{\theta} = \arg\max_\theta L(x_i|\theta)$.

In practice, $\hat{\theta}$ is found by deriving (wrt θ) the log-likelihood function

$$I(x_i|\theta) = \sum_{k=1}^{n} \ln\left(f(x_k|\theta)\right) \tag{4.7}$$

A MLE is asymptotically without bias and asymptotically normal.

→ N.B. : In order to compare the likelihood maximization and the method of moments, calculate both and use the one with the biggest likelihood.

The MLE for a discrete rv is $\hat{\theta} = \arg\{\max_\theta I(x_i|\theta)$, where $I(.)$ is the summ of log of the pmf :

$$I(x_i|\theta) = \sum_{i=1}^{n} \ln\left(p(x_k|\theta)\right) \tag{4.8}$$

Empirical mean and standard deviations - Properties

5.1 Properties of $\bar{X}$ and S^2

Let us consider n rv's $X_{i=1:n} \sim X$ and denote $\mu = \mathbb{E}(X)$ and $\sigma^2 = \mathbb{V}(X)$. The CLT states that whatever the distribution of X_i, the empirical mean tends to a normal :

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \qquad \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1) \tag{5.1}$$

The confidence interval for mu (or σ) at level $1-\alpha$ (often 5%) is an interval $[\mu_L, \mu_U]$ such that μ is in this interval with a probability $1-\alpha$.

$$P\left(z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \leq z_{1-\alpha/2}\right) = 1 - \alpha \qquad\qquad z_{\alpha/2} = -z_{1-\alpha/2} \tag{5.2}$$

As seen before, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is an unbiased estimator of $\mathbb{V}(X)$. If $X_i \sim N(\mu, \sigma^2)$, S^2 and $\bar{X}$ are independent and then

$$(n-1)\frac{S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \sim \chi^2_{n-1} \tag{5.3}$$

If $X \sim N(\mu, \sigma^2)$, S^2 is an estimator of σ. Since $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$, we infer that the $1-\alpha$ confidence interval for σ is

$$\left[\frac{n-1}{\chi^2_{n-1,1-\alpha/2}}S^2; \frac{n-1}{\chi^2_{n-1,\alpha/2}}S^2\right] \tag{5.4}$$

If $X_i \sim N(\mu, \sigma^2)$, then the following ratio is a Student's T rv : $\frac{\bar{X}-\mu}{\sqrt{S^2/n}} \sim t_{n-1}$. In that case, the confidence interval for μ is

$$\left[\bar{X} - t_{n-1,1-\alpha/2}\sqrt{S^2/n}; \bar{X} + t_{n-1,\alpha/2}\sqrt{S^2/n}\right] \tag{5.5}$$

5.1.1 Two populations

We consider two iid normal samples :

$$\begin{cases} X_1 = \{X_{1,1}, ..., X_{1,n_1}\} \sim N(\mu_1, \sigma_1^2) \Rightarrow \bar{X}_1 \sim N(\mu_1, \sigma_1^2/n_1) \\ X_2 = \{X_{2,1}, ..., X_{2,n_2}\} \sim N(\mu_2, \sigma_2^2) \Rightarrow \bar{X}_2 \sim N(\mu_2, \sigma_2^2/n_2) \end{cases}$$

By the properties of normal rv's,

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \implies \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \tag{5.6}$$

We define the unbiased estimators S_1^2, S_2^2 as before for the two populations and the following ratio is therefore a Fisher-Snedecor rv : $\frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} \sim F_{n_1-1,n_2-1}$.

If the two populations have the same variance, an unbiased "pooled" estimator of this variance is

$$S_{pool}^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \tag{5.7}$$

i.e. $\mathbb{E}(S_{pool}^2) = \sigma^2$ and $(n_1+n_2-2)S_{pool}^2/\sigma^2 \sim \chi^2_{n_1+n_2-2}$.

Furthermore, still with $\sigma_1 = \sigma_2$,

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_{pool}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{5.8}$$

Hypothesis testing - one population

A hypothesis is a claim about a parameter θ of a random sample. The null hypothesis is denoted by $H_0 = \theta \in \Theta_0$ and states the assertion to be tested. The alternative hypothesis is $H_1 = \theta \in \Theta_1$.

A statistical test is a procedure during which, based on some observations sampled from the probability distribution, we will make our decision of accepting or rejecting H_0.

	Reality	
Decisions	H_0 is true	H_1 is true
Accept H_0	Correct	Type 2 error (= false negative)
Reject H_0	Type 1 error (false positive)	Correct

The probability of making a type 1 error is called the significance level of the test and is usually denoted as $\alpha : P(\text{Type 1 error}) = \alpha$.
The decision to reject the null hypothesis is based on a test statistic $T(\cdot)$. Given a sample $\mathbf{X} = \{X_1, ...X_n\}$ of observations, $T(\mathbf{X}) : \mathbb{R}^n \to \mathbb{R}$. $T(\mathbf{X})$ is such that its distribution is known.

1. For a chosen α, we determine a rejection region $R_\alpha \subset \mathbb{R}$, such that $P(\text{Type 1 error}) = \alpha$.

2. We calculate $t = T(\mathbf{x})$ the observed value of $T(\mathbf{X})$.

3. Decision

   - If $t \in R_\alpha$, then reject H_0.
   - If $t \notin R_\alpha$, then do not reject H_0.

6.1 Single mean test

We cibsuder a iid sample $X_1, ..., X_n \sim N(\mu, \sigma^2)$ with unknown variance. We test $H_0 : \mu = \mu_0$ agaisnt three alternatives (with a different R_α for each) :

1. $H_1 : \mu > \mu_0$

2. $H_1 : \mu < \mu_0$

3. $H_1 : \mu \neq \mu_0$

A good choice for the test statistic is the Student's T : $T(\mathbf{X}) = \frac{\bar{X}-\mu_0}{\sqrt{S^2/n}} \sim t_{n-1}$.

If $t = T(\mathbf{X})$, the observed value of the Student's statistic, is "too far" from the mean of the student's distribution, it is likely that H_0 is false. We find a critical value c under the assumption that H_0 is true :

1. $P(T(\mathbf{X}) > c | H_0 \text{ is true}) = \alpha$
2. $P(T(\mathbf{X}) < c | H_0 \text{ is true}) = \alpha$
3. $P(T(\mathbf{X}) > c \text{ or } T(\mathbf{X}) < -c | H_0 \text{ is true}) = \alpha$

the critical values are the α, $1-\alpha$ or $(\alpha/2, 1-\alpha/2)$ quantiles of the Student's t.
We reject H_0 at the level α if $T(\mathbf{X}) = \sqrt{n}\frac{\bar{x}-\mu_0}{s}$

1. $H_1 : \mu > \mu_0 \qquad T(\mathbf{x}) > t_{n-1,1-\alpha}$
2. $H_1 : \mu < \mu_0 \qquad T(\mathbf{x}) < t_{n-1,\alpha}$
3. $H_1 : \mu \neq \mu_0 \qquad T(\mathbf{x}) < t_{n-1,\alpha/2}$ or $T(\mathbf{x}) > t_{n-1,1-\alpha/2}$

→ N.B. : if the variance σ^2 is known and the x_i are gaussian[1], then we use the test statistic $T(\mathbf{X}) = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim Z = N(0,1)$.

6.2 Single variance test

We consider a iid sample $X_1, ..., X_n \sim N(\mu, \sigma^2)$ with unknown variance. We test $H_0 : \sigma^2 = \sigma_0^2$ against three alternatives :

1. $H_1 : \sigma^2 \neq \sigma_0^2$
2. $H_1 : \sigma^2 > \sigma_0^2$
3. $H_1 : \sigma^2 < \sigma_0^2$

A good choice for the test statistic is the χ^2 test :

$$T(\mathbf{X}) = (n-1)\frac{S^2}{\sigma_0^2} \sim \chi^2_{n-1} \tag{6.1}$$

We work the same way we did for the single mean test and we find the following conditions :

1. $H_1 : \sigma^2 \neq \sigma_0^2 \qquad T(\mathbf{x}) < \chi_{n-1,\alpha/2}$ ou $T(\mathbf{x}) > \chi^2_{n-1,1-\alpha/2}$
2. $H_1 : \sigma^2 > \sigma_0^2 \qquad T(\mathbf{x}) > \chi_{n-1,1-\alpha}$
3. $H_1 : \sigma^2 < \sigma_0^2 \qquad T(\mathbf{x}) < \chi_{n-1,\alpha}$

[1] Gaussian means that they follow a normal distribution.

6.3 P-value

The P-value is the smallest level of significance for which the data indicate rejection of the null hypothesis.

Let $T(\mathbf{X})$ be a test statistic such that small values of T give evidence that H_0 is wrong. For a given sample $\mathbf{x}$, the p-value is :

$$p(\mathbf{x}) = P(T(\mathbf{X}) < T(\mathbf{x}) | H_0 \text{ is true}) \tag{6.2}$$

Let $T(\mathbf{X})$ be a test statistic such that high values of T give evidence that H_0 is wrong. For a given sample $\mathbf{x}$, the p-value is :

$$p(\mathbf{x}) = P(T(\mathbf{X}) > T(\mathbf{x}) | H_0 \text{ is true}) \tag{6.3}$$

Let $T(\mathbf{X})$ be a test statistic symmetric around zero such that high and small values of T give evidence that H_0 is wrong. For a given sample $\mathbf{x}$, the p-value is :

$$p(\mathbf{x}) = 2P(T(\mathbf{X}) > |T(\mathbf{x})| | H_0 \text{ is true}) \tag{6.4}$$

A small p-value indicates that H_0 is very unlikely. A high p-value informs us that H_0 is likely.

6.4 Comparison of two means

We consider two iid populations with the same variance :

$$\begin{cases} \mathbf{X}_1 = \{X_{1,1}, ..., X_{1,n_1}\} \sim N(\mu_1, \sigma^2) \Rightarrow \bar{X}_1 \sim N(\mu_1, \sigma^2/n_1) \\ \mathbf{X}_2 = \{X_{1,2}, ..., X_{1,n_2}\} \sim N(\mu_2, \sigma^2) \Rightarrow \bar{X}_2 \sim N(\mu_2, \sigma^2/n_2) \end{cases} \tag{6.5}$$

We test if the two samples have the same means, or more generically : $H_0 : \mu_1 - \mu_2 = \delta$, with δ a value that we choose, against

1. $H_1 : \mu_1 - \mu_2 > \delta$
2. $H_1 : \mu_1 - \mu_2 < \delta$
3. $H_1 : \mu_1 - \mu_2 \neq \delta$

6.4.1 The variance is known

If we remember the properties of a normal rv, $\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2})$. We therefore use the following test statistics :

$$T(\mathbf{X}_1, \mathbf{X}_2) = \frac{(\bar{X}_1 - \bar{X}_2) - \delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1) \tag{6.6}$$

We reject $H_0 : \mu_1 - \mu_2 = \delta$ at the level α if

1. $H_1 : \mu_1 - \mu_2 > \delta \qquad T(\mathbf{x}_1, \mathbf{x}_2) > z_{1-\alpha}$
2. $H_1 : \mu_1 - \mu_2 < \delta \qquad T(\mathbf{x}_1, \mathbf{x}_2) < z_{\alpha}$
3. $H_1 : \mu_1 - \mu_2 \neq \delta \qquad T(\mathbf{x}_1, \mathbf{x}_2) < z_{\alpha/2}$ or $T(\mathbf{x}_1, \mathbf{x}_2) > z_{1-\alpha/2}$

where z_α is the α-percentile of a $Z \sim N(0,1)$.

6.4.2 The variance is unknown

We use the test statistics

$$T(\mathbf{X}_1, \mathbf{X}_2) = \frac{(\bar{X}_1 - \bar{X}_2) - \delta}{S_{pool}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T_{n_1+n_2-2} \tag{6.7}$$

We reject $H_0 : \mu_1 - \mu_2 = \delta$ at the level α if

1. $H_1 : \mu_1 - \mu_2 > \delta \qquad T(\mathbf{x}_1, \mathbf{x}_2) > t_{n_1+n_2-2,1-\alpha}$
2. $H_1 : \mu_1 - \mu_2 < \delta \qquad T(\mathbf{x}_1, \mathbf{x}_2) < t_{n_1+n_2-2,\alpha}$
3. $H_1 : \mu_1 - \mu_2 \neq \delta \qquad T(\mathbf{x}_1, \mathbf{x}_2) < t_{n_1+n_2-2,\alpha/2}$ or $T(\mathbf{x}_1, \mathbf{x}_2) > t_{n_1+n_2-2,1-\alpha/2}$

where $t_{n_1+n_2-2,\alpha}$ is the α-percentile of a Student's T.

6.5 Comparison of two variances

We consider two iid populations with different variances :

$$\begin{cases} \mathbf{X}_1 = \{X_{1,1}, ..., X_{1,n_1}\} \sim N(\mu_1, \sigma_1^2) \\ \mathbf{X}_2 = \{X_{1,2}, ..., X_{1,n_2}\} \sim N(\mu_2, \sigma_2^2) \end{cases} \tag{6.8}$$

We test if the two samples have the same means, i.e. $H_0 : \sigma_1 = \sigma_2$ against

1. $H_1 : \sigma_1 \neq \sigma_2$
2. $H_1 : \sigma_1 > \sigma_2$
3. $H_1 : \sigma_1 < \sigma_2$

We use the test statistics

$$T(\mathbf{X}_1, \mathbf{X}_2) = \frac{S_1^2}{S_2^2} \sim F_{n_1-1,n_2-1} \tag{6.9}$$

with F Fisher's test. We reject $H_0 : \sigma_1 = \sigma_2$ at the level α if

1. $H_1 : \sigma_1 \neq \sigma_2 \qquad T(\mathbf{x}_1, \mathbf{x}_2) < F_{n_1-1,n_2-1,\alpha/2}$ or $T(\mathbf{x}_1, \mathbf{x}_2) > F_{n_1-1,n_2-1,1-\alpha/2}$
2. $H_1 : \sigma_1 > \sigma_2 \qquad T(\mathbf{x}_1, \mathbf{x}_2) > F_{n_1-1,n_2-1,1-\alpha}$
3. $H_1 : \sigma_1 < \sigma_2 \qquad T(\mathbf{x}_1, \mathbf{x}_2) < t_{n_1-1,n_2-1,\alpha}$

where $F_{n_1-1,n_2-1,\alpha}$ is the α-percentile of a Fisher.

Linear regression

We observe n realizations Y_i that is related to k factors $(x_{i,1}, ..., x_{i,k})^T$ for $i = 1, ..., n$. We postulate the following linear relation :

$$Y_i = \beta_0 + \beta_1 x_{i,1} + ... + \beta_k x_{i,k} + \epsilon_i \qquad i = 1, ..., n \tag{7.1}$$

where $\epsilon_i \sim N(0, \sigma^2$ is the noise of the system. In matrix notations, we have

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k} \end{pmatrix} \qquad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \tag{7.2}$$

$\mathbf{Y} and \epsilon$ are n vectors, β is a $k + 1$ vector and X is a $N \times (k + 1)$ matrix. We call the β_j the regression coefficients.
We can reformulate the equation Equation 7.1 as $\mathbf{Y} = \mathbf{X}\beta + \epsilon$.

The best $\hat{Y}$ prediction of Y for a given value of $\mathbf{x} = (1, x_1, ..., x_k)^T$ is

$$\hat{Y} = \mathbb{E}(Y|\mathbf{x}) = \mathbf{x}^T \beta \tag{7.3}$$

However, β, σ^2 are unknown. We denote by $\hat{beta}$ and $\hat{\sigma}^2$ their estimates, whose values we determine by a method of estimation, such as the likelihood maximization.

With this method of estimation, we find that, for $\theta = (\beta, \sigma)$,

$$I(\theta) \propto -\sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \tag{7.4}$$

which is the least squares equation.

7.1 Least square minimization

The estimate $\hat{\beta}$ of β minimizes the sum of squared errors (SSE) :

$$\hat{\beta} = \arg_\beta \min \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \Longrightarrow \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \tag{7.5}$$

Furthermore, the best prediction of Y_i for a vector of factor $\mathbf{x}_i$ is

$$\hat{\mathbf{y}}_i = \mathbf{x}_i^T \hat{\beta} \qquad i = 1, ..., n \Longleftrightarrow \hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{H}y \tag{7.6}$$

with $\hat{y} = (y_1, ..., y_n)^T$ and $H = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ the hat matrix[1].

[1] The hat matrix is symmetric ($\mathbf{H} = \mathbf{H}^T$) and idempotent ($\mathbf{HH} = \mathbf{H}$).

→ N.B. : these values for the estimators are correct only under the hypothesis that the data are gaussian. If it is not the case, they have different values, but we can still use the least squares minimization.

7.1.1 Simple linear regression

In a simple regression, we only have one explanatory factor ($k = 1$) and the equation is $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Then

$$\{\beta_0, \beta_1\} = \arg_\beta \min \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{7.7}$$

Of which we can derive

$$\begin{cases} \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases} \tag{7.8}$$

We can see that $\hat{\beta}_1$ is proportional to the correlation coefficient between X and Y.

7.1.2 Goodness of fit

Let us note $\hat{y}_i = \mathbf{x}_i^{\hat{T}} \hat{\beta}$. Then

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \iff SSTotal = SSError + SSRegression \tag{7.9}$$

Where $SST = (n-1)S_Y^2$, SSE is the noise, and SSR is proportional to the variance explained by the model.

Let us defined $R^2 \in [0, 1]$ the proportion of the variance explained by the model : $R^2 = \frac{SSR}{SSR} \iff 1 - R^2 = \frac{SSE}{SST}$. The closer R^2 is to unity, the better is the model.

7.2 Properties of regression coefficients

- $\hat{\beta}$ is an unbiased Gaussian estimator of β, i.e. $\mathbb{E}(\hat{\beta}) = \beta : \hat{\beta} \sim N\left(\beta, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}\right)$
- An unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n - (k+1)} (\mathbf{Y} - \mathbf{X}\hat{\beta})^T (\mathbf{Y} - \mathbf{X}\hat{\beta}) = \frac{SSE}{n - (k+1)} \implies (n - (k+1)) \frac{\hat{\sigma}^2}{\sigma^2} = \frac{SSE}{\sigma^2} \sim \chi^2_{n-(k+1)} \tag{7.10}$$

which is a chi-square variable with $n - (k+1)$ degrees of freedom.

7.3 Simple linear regression

$\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators. If $\bar{x^2} = \frac{1}{n} \sum_{i=1}^n x_i^2$, their variances are

$$\mathbb{V}(\hat{\beta}_0) = \frac{\sigma^2 \bar{x^2}}{S_{xx}} \qquad \mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} \qquad \mathbb{C}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2 \bar{x}}{S_{xx}} \tag{7.11}$$

where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$. Since $\epsilon \sim N(0, \sigma^2)$,

$$\hat{\beta}_0 \sim N(\beta_0, \frac{\sigma^2 \bar{x}^2}{S_{xx}} \qquad \hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}} \qquad (n-2)\frac{\hat{\sigma}^2}{\sigma^2} = \frac{SSE}{\sigma^2} \sim \chi^2_{n-2} \tag{7.12}$$

In practice, σ^2 is unknown and we replace it by $\hat{\sigma}^2$. In consequence, $\hat{\beta}_0$ and $\hat{\beta}_1$ become Student's T.

7.4 Test of the significance of linear regression

From properties of S^2, if $\beta_1 = ... = \beta_k = 0$, the normalized SST is a chi-square rv with $n-1$ degrees of freedom.

$$\frac{SST}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sim \chi^2_{n-1} \tag{7.13}$$

and the normalized SSE is a chi-square rv with $n-(k+1)$ degrees of freedom.

$$\frac{SSE}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \sim \chi^2_{n-(k+1)} \tag{7.14}$$

Since $SST = SSE + SSR$ and as a χ^2_n rv is a sum of n rv's,

$$\frac{SSR}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \sim \chi^2_k \tag{7.15}$$

Therefore, if $\beta_1 = ... = \beta_k = 0$, the next ratio is a Fisher rv :

$$F^* = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-(k+1))} \sim F_{k,n-(k+1)} \tag{7.16}$$

If F^* is "too small", then the assumption of linearity between Y and $\mathbf{X}$ must be rejected.

The significance test of regression is $\begin{cases} H_0 : \beta_1 = ... = \beta_k = 0 \\ H_1 : \beta_j \neq 0 \text{ for some } j \in \{1, ..., k\} \end{cases}$, with the test statistics F^* above.
Reject H_0 at a confidence level α

- if $F^* > F_{k,n-(k+1),1-\alpha}$;
- if the p-value $p_{val} = P(F^* < F_{k,n-(k+1)})$ is lower than α.

7.4.1 Tests of regression coefficients

Let $c_{j,j}$ be the j^{th} diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$. Estimators $\hat{\beta}$ of linear regression coeffients are Student's T rv :

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{c_j j}} \sim t_{n-(k+1)} \tag{7.17}$$

We use this result to test the significance of each β_j : $H_0 : \beta_j = \beta_{j,0}$ against three alternatives :

1. $H_1 : \beta_j > \beta_{j,0}$
2. $H_1 : \beta_j < \beta_{j,0}$
3. $H_1 : \beta_j \neq \beta_{j,0}$

with the test statistics $T_j^* = \frac{\hat{\beta}_j - \beta_{j,0}}{\hat{\sigma}\sqrt{c_{jj}}} \sim t_{n-(k+1)}$. We reject H_0 at the level α if

1. $H_1 : \beta_j > \beta_{j,0} \qquad T_j^* > t_{n-(k+1),1-\alpha}$
2. $H_1 : \beta_j < \beta_{j,0} \qquad T_j^* < t_{n-(k+1),\alpha}$
3. $H_1 : \beta_j \neq \beta_{j,0} \qquad T_j^* < t_{n-(k+1),\alpha/2}$ or $T_j^* > t_{n-(k+1),1-\alpha/2}$

And we find a $1-\alpha$ confidence interval as follows :

$$\beta_j \in \left[\hat{\beta}_j + t_{n-(k+1),\alpha/2}\hat{\sigma}\sqrt{c_j j}; \hat{\beta}_j + t_{n-(k+1),1-\alpha/2}\hat{\sigma}\sqrt{c_j j}\right] \tag{7.18}$$

or, as the Student's T is symmetric,

$$\beta_j \in \left[\hat{\beta}_j - t_{n-(k+1),1-\alpha/2}\hat{\sigma}\sqrt{c_j j}; \hat{\beta}_j + t_{n-(k+1),1-\alpha/2}\hat{\sigma}\sqrt{c_j j}\right] \tag{7.19}$$

In case of simple linear regression , we have $c_{0,0} = \frac{\bar{x}^2}{S_{xx}}$ and $c_{1,1} = \frac{1}{S_{xx}}$ and the confidence interval is

$$\beta_1 \in [\hat{\beta}_1 - t_{n-(k+1),1-\alpha/2}\hat{\sigma}\sqrt{S_{xx}^{-1}}; \hat{\beta}_1 + t_{n-(k+1),1-\alpha/2}\hat{\sigma}\sqrt{S_{xx}^{-1}}] \tag{7.20}$$

For β_0, we test $H_0 : \beta_0 = \beta_{0,0}$ against $H_1 : \beta_0 \neq \beta_{0,0}$ with the test statistics $T_0^* = \frac{\hat{\beta}_0 - \beta_{0,0}}{\hat{\sigma}\sqrt{\bar{x}^2 S_{xx}^{-1}}} \sim t_{n-2}$. The confidence interval is

$$\beta_0 \in [\hat{\beta}_0 - t_{n-(k+1),1-\alpha/2}\hat{\sigma}\sqrt{\bar{x}^2 S_{xx}^{-1}}; \hat{\beta}_1 + t_{n-(k+1),1-\alpha/2}\hat{\sigma}\sqrt{\bar{x}^2 S_{xx}^{-1}}] \tag{7.21}$$

7.4.2 Prediction interval

For a model $Y = \beta_0 + \beta_1 X + \epsilon$, the prediction for an unobserved value $X = x_0$ is $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. The prediction interval for Y_0 at level α is provided by

$$[\hat{y}_0 - S_{pred} t_{n-2,1-\alpha/2}; \hat{y}_0 + S_{pred} t_{n-2,1-\alpha/2}] \tag{7.22}$$

where $S_{pred}^2 = \hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$

7.5 Analysis of Variance (ANOVA)

ANOVA is used to compare the means of n different sets of data. To compare these, we will set $n-1$ binary (or dummy) variables

$$X_i = \begin{cases} 1 \text{ if the data is from set } i \\ 0 \text{ otherwise} \end{cases} \tag{7.23}$$

We now have $Y = \beta_0 + \sum_i \beta_i X_i + \epsilon$ and our hypothesis is $\begin{cases} H_0 : \beta_i = 0 & \forall i \\ H_1 : \exists i, & \beta_i \neq 0 \end{cases}$

Data science - Supervised learning

8.1 Introduction

- We write a set of indexed elements as : $\{a_i\}_{i=1}^N \equiv \{a_i : 1 \leq i \leq N\}$.
- Most of the time, the elements are in a p-dimensional space : $a_i \in \mathbb{R}^p$.
- The letter N is always the size of a (data)set.
- Matrices and vectors with N rows are written in bold : $\mathbf{X}, \mathbf{u}, \dots$
- We write $[N] \coloneqq \{1, \dots, N\}$.
- $I(a \neq b) = 1$ if $a \neq b$ and 0 otherwise.

8.1.1 Regression vs Classification

The most common supervised learning tasks are regression and classification. Both are doing predictions, but on different objects.

- Regressions predict ordered values (=quantitative), continuous or discrete.
- Classifications predict classes or categories (= qualitative).

→ N.B.: for binary data, there is no distinction between regression and classification.

To turn classification into a regression, we use dummy variables: from the classes $\mathcal{G} = \{g_1, \dots, g_k\}$, we create the vectors $e_k \coloneqq (0, \dots, 0, 1, 0, \dots, 0)^T \in \mathbb{R}^K$, where the 1 is the k-th component.

8.2 Modeling

We have the features (=inputs) $X = (x_1, \dots, x_p)^T \in \mathbb{R}^p$ and the outcome (=labels) Y, quantitative or qualitative. We assume that X and Y are related. This means that there exists a function $f : \mathbb{R}^p \to \mathbb{R}$ such that

- For a regression, $Y = f(X) + \varepsilon$, for some rv ε, the noise.
- For a binary classification, $P(Y = f(X)) = 1 - \eta$ and $P(Y \neq f(X)) = \eta$, for a misclassification error $0 \leq \eta \leq 1$.

The objective is to find a reliable estimate $\hat{f}$ of f for doing predictions of (a sampling of) Y given (a sampling of) X.

- Training stage:
  1. Get a representative data set $\mathcal{T}$ of your task, i.e. a sampling of (X, Y), a set of N measured pairs of {"input","label"}: $\mathcal{T} := \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^p \times \mathbb{R}$.
  2. Clean, preprocess and transform the data.
  3. Train/fit a machine learning model, i.e. find the model parameters.
- Prediction stage:
  1. Clean/preprocess/transform the new/independent data.
  2. Apply the model on them to make predictions.

8.3 Multivariate linear modeling

8.3.1 Definition

Given a vector of inputs $X = (X_1, \dots, X_p)^T$, we want to predict the output Y, assuming the linear model

$$\hat{Y} = f_{\hat{\beta}}(X) := \sum_{j=1}^{p} X_j \hat{\beta}_j = \mathbf{X}^T \hat{\beta} \tag{8.1}$$

with the coefficients $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^T$.
This model is linear; bias free, i.e. $\hat{\beta}_0 = 0$; and global/parametric.

Geometrically, the problem is to fit a hyperplane through the data ($\equiv$ finding an optimal $\hat{\beta}$).

8.3.2 Fitting

Given the training set $\mathcal{T} := \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^p \times \mathbb{R}$, we pick $\hat{\beta}$ as the minimizer of the empirical risk

$$\hat{R}(\beta) := \frac{1}{N} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 = \frac{1}{N} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \tag{8.2}$$

The solution to this problem is given by $\hat{\beta} := (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, and we have the predictions $\hat{y} = \mathbf{X}\hat{\beta} = \mathbf{H}\mathbf{y}$, with $\mathbf{H} := \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, the hat matrix.

→ N.B.: for this solution to exist, we need $N > p$ and $\mathbf{X}^T\mathbf{X}$ invertible.

8.4 Classification with a linear model

Let $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^2 \times \{0, 1\}$, with the classes

1. label $y_i = 1$: $x_i \sim N(\mu_1, \sigma^2\mathbf{I})$
2. label $y_i = 0$: $x_i \sim N(\mu_2, \sigma^2\mathbf{I})$

→ N.B.: $N(\mu, \sigma^2\mathbf{I}) \sim c_\sigma \exp\left(-\frac{\|x-\mu\|^2}{2\sigma^2}\right)$, with $x, \mu \in \mathbb{R}^2$, $\sigma > 0$ and $c_\sigma := (2\pi\sigma^2)^{-p/2}$.

8.4.1 LS regression

Least-square regression (LS) provides a continuous $\hat{Y}$, but $Y \in \{0,1\}$ in our example. We define a threshold $\tau > 0$ and, for $x \in \mathbb{R}^2$, the estimated class $\hat{G}$ of x is

$$\hat{G}(x) = \mathcal{B}_\tau(\mathbf{x}^T\hat{\beta}) := \begin{cases} 1 \text{ if } (\mathbf{x}^T\hat{\beta}) > \tau \\ 0 \text{ if } (\mathbf{x}^T\hat{\beta}) \leq \tau \end{cases} \tag{8.3}$$

Thus, a LS classifier is a LS regression combined to a binary conversion $\mathcal{B}_\tau$.

8.4.2 k-Nearest Neighbors Method

Given a number of neighbors $k \in \mathbb{N}$, we first define $N_k(x)$ the set of j closest points to x in $\{x_i\}_{i=1}^N \subset \mathbb{R}^p$.
In k-Nearest Neighbors regression, the value $\hat{Y}(x)$ assigned to x is the average of the points' label in $N_k(x)$.

$$\hat{Y}(x) = \frac{1}{k}\sum_i y_i \tag{8.4}$$

This method works with a local average and is non-parametric.

A k-NN classifier is a k-NN regression combined to a binary conversion $\mathcal{B}_\tau$.
However, this classifier is not always good, as the error on the test set isn't monotonously decreasing with N/k increasing.

In this method, there are regions in which we can move the points it contains without changing the solution. Theses zones are the parameters of the method and there are N/k of them.

	Assumptions	Accuracy	Stability	Number of parameters	Time complexity
LS	Strong (linear)	Low	High	p	Large
k-NN	Mild	Low/high	High/Low	$\approx N/k$	Low

8.5 Models for Supervised Learning

The SL objective is using an algorithm $\mathcal{A}$ to learn a prediction function $\hat{f}$ from a dataset $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^N$. This means that $\hat{Y} = \hat{f}(X)$ predicts $Y = f(X)$ if $\mathcal{T}$ samples well (X, Y). It is however impossible to solve this problem without assumptions; we assume that $f \in \mathcal{F}$, with $\mathcal{F}$ a restricted class of functions. There are two types of function class $\mathcal{F}$:

- parametric: $\mathcal{F} := \{f_\beta : \beta \in \mathcal{B}\}$, with β in the parameter set $\mathcal{B}$.
- nonparametric: $\mathcal{F}$ can be non explicitly described (e.g. k-NN).

In both cases, $\mathcal{F}$ depends on hyper-parameters γ.

8.5.1 Parametric models

Let us assume a parametric model with parameters $\beta = (\beta_0, \dots, \beta_)^T$. Learning the function $f \coloneqq \{f_\beta(X) \coloneqq \beta^T \varphi(X) : \beta \in \mathbb{R}^{p+1}\}$ is minimizing a cost function l over $\mathcal{T}$, e.g. the squared loss.

$\rightarrow$ N.B.:p is a hyperparameter.

These models are easy to optimize and are applicable to any dimensions, but we need to avoid overfitting and a good feature vector φ for good parameters.

8.5.2 Non-parametric models

In those models, we make no explicit assumption on the function form of f, but implicit assumption. For example, in k-NN, we assume that $\mathcal{F}$ is the piecewise constant functions.
These models can fit a larger set of unknown functions f and are very flexible, but we need much more data and are less interpretable.

8.5.3 Model accuracy

A good model $\hat{f}$ is such that, on a new input x with label y, we get a small predicition error, i.e. $\hat{f}(x) \approx y$. We define the risk for $Y = f(X) + \varepsilon$:

$$R(\hat{f}) \coloneqq \mathbb{E}\left((Y - \hat{f}(X))^2\right) = \left(f(X) - \hat{f}(X)\right)^2 + \mathbb{V}(\varepsilon) \tag{8.5}$$

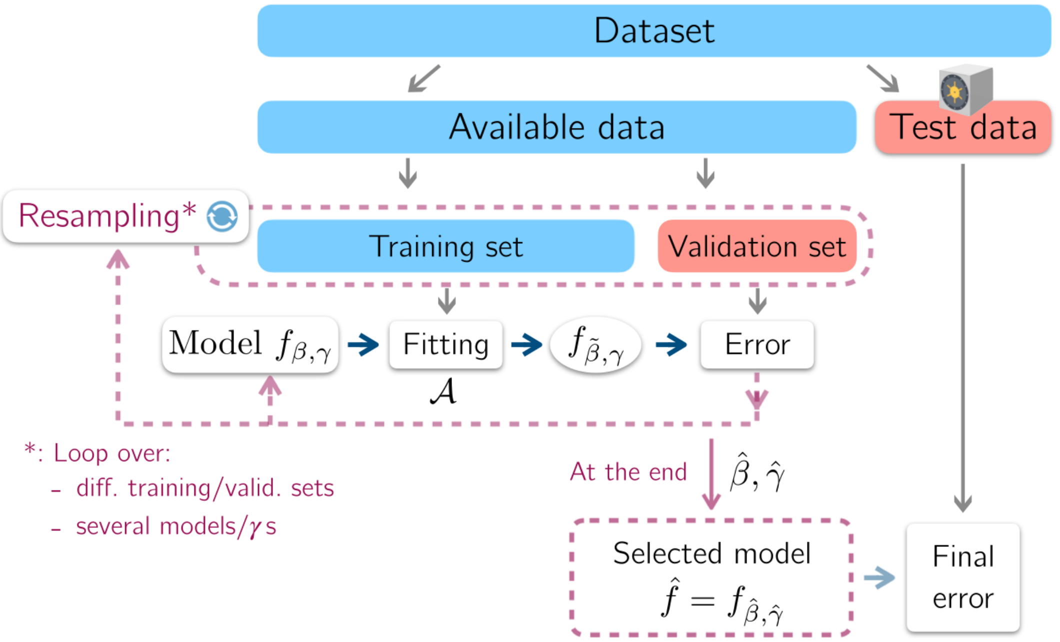
The only reducible term is the first one, as we cannot influence ε, the noise.

To assess the accuracy in practice, we have two tools:

- Mean Squared Error (MSE) for regressions: $MSE = 1/N \sum_{i=1}^{N} \left(y_i - \hat{f}(x_i)\right)^2$.
- Misclassification error rate (Err) for classifications: $Err = 1/N \sum_{i=1}^{N} I(y_i \neq \hat{f}(x_i))$.
- Both are empirical risks, i.e. obtained from $\hat{R}(\hat{f}) \coloneqq 1/N \sum_{i=1}^{N} l(y_i, \hat{f}(x_i))$.

8.6 Resampling methods

Resampling methods are repeatedly drawing samples (at random) from a training set, and refitting a model of interest on the set of these samples. This is done for model assessment, i.e. evaluate the model's performance, and for model selection, i.e. select (hyper-)parameters or improve the feature space.



8.6.1 Validation set approach

The ideal objective in supervised learning is to reduce the error of a machine learning method on new observations, i.e. the test error. However, new observations (= test dataset) are often not available. We can then hold out a subset of the training set from the fitting process and apply the fitted machine learning method (MLM) to the held out data. To do this, we randomly divide the available data into a training set and a validation set, on which we estimate the test error. Mathematically, this means that, to have a 50%/50% split,

1. Given a number N of observations, a model $\hat{f}$, and a random permutation function $\pi : [N] \to [N]$, we form
   - a training set $\mathcal{T}_{tr} = \{(x_{pi(i)}, y_{pi(i)})\}_{i=1}^{N/2}$
   - a validation set $\mathcal{T}_v = \{(x_{pi(i)}, y_{pi(i)})\}_{i=N/2+1}^{N}$
2. We fit $\hat{f}$ on $\mathcal{T}_r$ and compute the validation error (MSE or Err).

The drawbacks of this method are that the estimate of the test error rate is variable, and the training set uses only half of the observations.

8.6.2 Leave-one-out cross-validation

The LOOCV approach is to use a training set of size $N - 1$ and a validation set of size 1. There are thus N possible configurations and we use them all. Mathematically, for each $i \in \{1, \dots, N\}$,

1. Define $\mathcal{T}_i := \{(x_j, y_j) : j \in [N], j \neq i\}$ for $1 \leq i \leq N$.

2. Fit the model on $\mathcal{T}_i$ and get $\hat{f}_i$.
3. Compute the prediction error (MSE or Err).

In LOOCV, the test error estimation is an average of all the prediction errors:

$$\begin{cases} CV_{(N)} = \frac{1}{N}\sum_{i=1}^{N} MSE_i \\ CV_{(N)} = \frac{1}{N}\sum_{i=1}^{N} Err_i \end{cases} \tag{8.6}$$

This method provides a better test error estimate and it is easier to detect optimal parameters, but it is very time-consuming.

8.6.3 K-Fold Cross Validation

The K-Fold CV is the same method as LOOCV, but with a validation set of size K instead of 1. It is mathematically very similar to LOOCV:

1. Create the K folds: $Fold_i \subset [N]$, with $|Fold_i| = N/K$, $\cup_{i=1}^{K} Fold_i = [N]$.
2. For each split $i \in \{1, \dots, K\}$, fit the model on $\mathcal{T}_i := \{(x_j, y_j) : j \notin Fold_i\}$ and get $\hat{f}_i$; then compute the prediction error on i-th fold $\mathcal{T}_i$.

In KFCV, the test error estimate is an avergae on all splits:

$$\begin{cases} CV_{(K)} = \frac{1}{K}\sum_{i=1}^{K} MSE_i \\ CV_{(K)} = \frac{1}{K}\sum_{i=1}^{K} Err_i \end{cases} \tag{8.7}$$

It is very similar to LOOCV, but is much faster to compute.

8.7 Classification methods

An algorithm performing classification is called a classifier, and it often predicts the probabilities that the outcome is in each category. A decision is made from these probabilities afterwards.

8.7.1 Linear classification methods

Given the Q labels $\mathcal{G} = \{g_1, \dots, g_Q\}$ and an input space $\mathbb{R}^p$, we classify the data by dividing $\mathbb{R}^p$ in Q regions, delineated by linear boundaries. The q-th boundary is $\{x : \delta_q(x) = 0\}$, i.e. a plane, given a discriminant function $\delta_q(x)$.

Logistic regression in 1D

Let be a binary class problem in 1D: $Y = G \in \{g_1 = 0, g_2 = 1\}$, with $X \in \mathbb{R}$. The logistic regression models the probability that $Y = 1$ or 0 given $X = x$, i.e. $h(x) := p(Y = 1|X = x)$. we use the logistic function, with $\beta = (\beta_0, \beta_1)^T$

$$h(X) = h_\beta(X) := S(\beta_0 + \beta_1 X) := \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} \tag{8.8}$$

$\rightarrow$ N.B.: $p(Y = 0|X = x) = 1 - h(x)$.

Note that the logistic regression is not a classifier, because $p(Y|X) \in [0,1]$. The logistic classifier would be the logistic regression coupled to a binary conversion. Given a probability threshold $0 \leq p^* \leq 1$, the estimated class $\hat{G}(x)$ of $x \in \mathbb{R}$ reads

$$\hat{G}(x) = \begin{cases} 1 \text{ if } p(Y=1|X=x) \geq p^* \\ 0 \text{ otherwise} \end{cases} = \begin{cases} 1 \text{ if } \log \frac{p(Y=1|X=x)}{p(Y=0|X=x}) \geq \tau^* := \log \frac{p^*}{1-p^*} \\ 0 \text{ otherwise} \end{cases} \tag{8.9}$$

$$= \begin{cases} 1 \text{ if } \beta_0 + \beta_1 x \geq \tau^* \\ 0 \text{ otherwise} \end{cases} \tag{8.10}$$

$\hat{G}$ splits $\mathbb{R}$ according to the sign of $\delta(x) := \beta_0 + \beta_1 x - \tau^*$.

For p inputs $X_1, \ldots, X_p \in \mathbb{R}$ instead of one $X \in \mathbb{R}$, we generalize the model to

$$X = (X_0 = 1, X_1, \ldots, x_p)^T \in \mathbb{R}^{p+1} \qquad \beta = (\beta_0, \beta_1, \ldots, \beta_p)^T \in \mathbb{R}^{p+1} \tag{8.11}$$

so that

$$h(X) = h_\beta(X) := \frac{\exp\left(\sum_{i=0}^p \beta_i X_i\right)}{1 + \exp\left(\sum_{i=0}^p \beta_i X_i\right)} = \frac{\exp\left(\beta^T X\right)}{1 + \exp\left(\beta^T X\right)} = S(\beta^T X) \tag{8.12}$$

We call β_0 the intercept and $(\beta_1, \ldots, \beta_p)^T$ the direction in $\mathbb{R}^p$.

Fitting the logistic regression model

1. $Y|X$ is binary and $p(Y=1|X=x) = h_\beta(x) \implies Y|X \sim Ber(h_\beta(x))$.
2. Under the logistic regression model, likelihood of $Y = y$ given $X = x$ is

$$p(Y=y|X=x) = \begin{cases} h_\beta(x) \text{ if } y = 1 \\ 1 - h_\beta(x) \text{ if } y = 0 \end{cases} = h_\beta(x)^y (1 - h_\beta(x))^{1-y} \tag{8.13}$$

3. Given the dataset $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^N$, the likelihood of $\mathcal{T}$ is:

$$L(y_1, \ldots, y_N | \beta, x_1, \ldots, x_n) := p(Y_1 y =, \ldots, Y_N = y_N | X_1 = x_1, \ldots, X_N = x_N)$$

$$= \prod_{i=1}^N h_\beta(x_i)^{y_i} (1 - h_\beta(x_i))^{1-y_i} \tag{8.14}$$

Maximizing L is equivalent to minimizing the negative log-likelihood $\mathcal{L}(\beta)$, i.e.

$$\mathcal{L}(\beta) = \mathcal{L}(\beta, \mathcal{T}) := -\log L(y_1, \ldots, y_N | \beta, x_1, \ldots, x_N) \tag{8.15}$$

$$:= -\left(\sum_{i=1}^N y_i \log h_\beta(x_i) + (1 - y_i) \log(1 - h_\beta(x_i)) \right) \tag{8.16}$$

$$:= \sum_{i=1}^N \log(1 + e^{\beta^T x_i}) - y_i \beta^T x_i \tag{8.17}$$

Therefore, the optimal logistic parameters $\hat{\beta}$ is obtained from

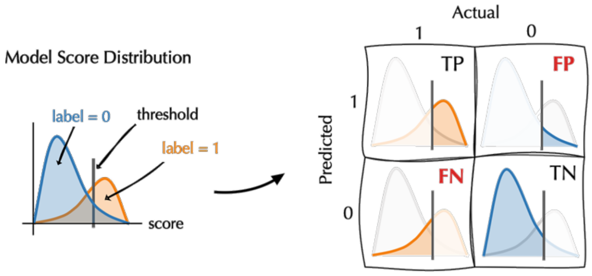
$$\hat{\beta} := \arg\min_\beta \mathcal{L}(\beta) = \sum_{i=1}^N \log(1 + e^{\beta^T x_i}) - y_i \beta^T x_i \tag{8.18}$$

Performance analysis - Accuracy

$$Acc^{emp}(\hat{G}) := \frac{1}{N}\sum_{i=1}^{N} I(y_i = \hat{G}(x_i)) = 1 - \frac{1}{N}\sum_{i=1}^{N} I(y_i \neq \hat{G}(x_i)) = 1 - Err \tag{8.19}$$

Performance analysis - TP/TN/FP/FN

Given the output $\hat{G}$, we define the True Positive, False Negative, False Positive and True Negative. In general, for a Q-class classification problem, these numbers are computed by the confusion matrix $\mathcal{C} \in \mathbb{N}^{Q\times Q}$, with $\mathcal{C}_{ij}$ the number of elements of class i classified in class j. In a binary class,



$$\mathcal{C} = \begin{pmatrix} TN & FP \\ FN & TP \end{pmatrix} \tag{8.20}$$

Performance analysis : Confusion matrix

The confusion matrix is computed from the training set :

- Use K-Fold CV, predict all y_i using the $K-1$ folds $\not\ni i \implies$ this gives all $\hat{y}_i$.
- use all the true/predicted label pairs $\{(y_i, \hat{y}_i)\}_{i=1}^N$ to compute $\mathcal{C}$.

Performance analysis - Precision and Recall

In binary classification, the precision and recall are defined as follows :

$$Pre^{emp}(\hat{G}) = \frac{TP}{TP+FP} \approx \frac{P(Y=1\&\hat{G}(X)=1)}{P(\hat{G}(X)=1)} \tag{8.21}$$

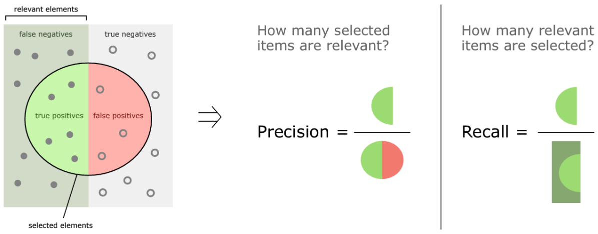
It is the amount of true positives over perceived positives.

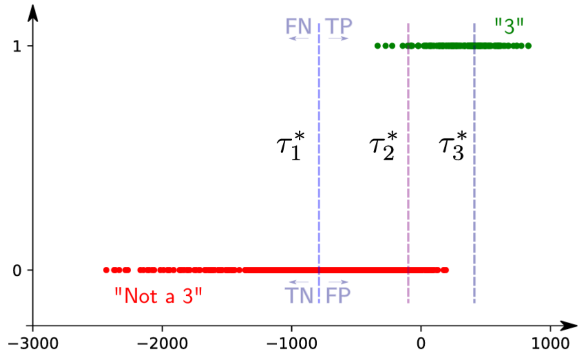
$$Rec^{emp}(\hat{G}) = \frac{TP}{TP+FN} \approx \frac{P(Y=1\&\hat{G}(X)=1)}{P(Y=1)} \tag{8.22}$$

It is the amount of true positives over the real positives. We can now define the F1 score, that is the harmonic mean of precision and recall :

$$F_1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} \in [0,1] \tag{8.23}$$

We can choose the value of the parameter τ^* depending on the level of precision and recall we want:





Receiver Operating Curve (ROC)

The ROc is the plot of TPR (recall) vs FPR (1-specificity), with :

$$TPR := \frac{TP}{P} = \frac{TP}{TP+FN} \qquad FPR := \frac{FP}{N} = \frac{FP}{FP+TN} \tag{8.24}$$

The AUC is the area under the ROC. A good classifier has a AUC near 1[1].

8.8 Statistical Decision Theory

The objective of SDT is to explain the behaviors of models on unseen data, i.e. that is not in the initial dataset.

8.8.1 Context

Let $X \in \mathbb{R}^p$ be a real-valued random input vector and $Y \in \mathcal{Y}$, with $\mathcal{Y} = \mathbb{R}$ for regressions and $\mathcal{Y} = \mathcal{G}$ for classifications, be a random output value.

Both are liked by a joint distribution $p(X, Y)$ over $\mathbb{R}^p \times \mathcal{Y}$. The training dataset is the joint sampling of (X, Y) :

$$\mathcal{T} := \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^p \times \mathcal{Y} \qquad (x_i, y_i) \sim_{i.d.d.} X \times Y \tag{8.25}$$

Our objective here is to find a function f to predict Y, given X, i.e. $Y \approx \hat{Y} := f(X)$. In this model, we find f by minimizing a prediction error defined by a loss function:

$$l : (Y, f(X)) \in \mathcal{Y} \times \mathcal{Y} \to l(Y, f(X)) \in \mathbb{R}_+ \tag{8.26}$$

such that

$$Y \approx f(X) \iff l(Y, f(X)) \approx 0 \tag{8.27}$$

[1] A random classifier has a AUC of 0.5.

8.8.2 Risk

Given a loss function l and decision f, we define the expected risk :

$$R(f) := \mathbb{E}l(Y, f(X)) = \int_{\mathcal{Y}} \int_{\mathbb{R}^p} l(y, f(x)) p(x, y) dx dy \tag{8.28}$$

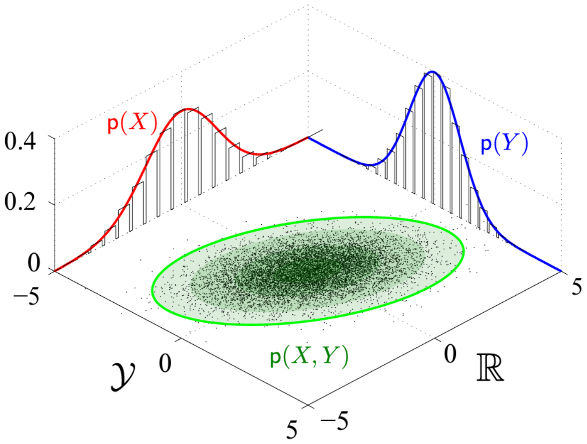
The risk is defined as the expected error of the model, and the general objective is to minimize it.

We also define the empirical risk:

$$\hat{R}(f) := \frac{1}{N} \sum_{i=1}^{N} l(y_i, \hat{f}(x_i)) \tag{8.29}$$

→ N.B.: if $N \to \infty$, $\hat{R}(f) \to R(f)$.

8.8.3 Bayes risk and predictor



For $p(X, Y) = p(Y, X)$, the marginals $p(X)$ and $p(Y)$ are

$$p(X = x) := \int_{\mathcal{Y}} p(X = x | Y = y) dy \tag{8.30}$$

$$p(Y = y) := \int_{\mathbb{R}^p} p(Y = y | X = x) dx \tag{8.31}$$

$$\tag{8.32}$$

and the pdf of Y conditioned on X is

$$p(Y|X) := \frac{p(Y, X)}{p(X)} \leq 1 \iff p(Y, X) = p(Y|X)p(X) \tag{8.33}$$

Accordingly, Bayes' theorem reads

$$p(Y|X) = p(X|Y) \frac{p(Y)}{p(X)} \tag{8.34}$$

Additionnally, the law of total expectation (LTE) is the following: for any rv's U, V, and a function $g(U, V)$,

$$\mathbb{E}g(U, V) = \mathbb{E}_{U,V}g(U, V) = \mathbb{E}_U(\mathbb{E}_{V|U}(g(u, V)|U = u)) \tag{8.35}$$

where

$$\mathbb{E}_{V|U}(g(u, V)|U = u) := \int g(u, v)p(V = v|U = u)dv \tag{8.36}$$

If we apply the LTE to the risk,

$$R(f) = \mathbb{E}l(Y, f(X)) = \mathbb{E}_X\mathbb{E}_{Y|X}(l(Y, f(X))|X = x) \tag{8.37}$$

where we define the conditional risk, depending on x and $p(Y|X)$:

$$r : z \in \mathbb{R} \to r(z|X = x) := \mathbb{E}_{Y|X}(l(Y, z)|X = x) \tag{8.38}$$

Finding the optimal function f is thus equivalent to find f such that for any $x \in \mathbb{R}^p$, we get minimum conditional risk $r(f(x)|X = x)$.

If we know $p(Y|X)$, the Bayes predictor is

$$f^*(x) := \arg\min_{z\in\mathcal{Y}} r(z|X = x) = \arg\min_{z\in\mathcal{Y}} \mathbb{E}_{Y|X}(l(Y, z)|X = x) \tag{8.39}$$

and its risk, the Bayes risk, is

$$R^* := R(f^*) = \mathbb{E}_X \min_{z\in\mathbb{Y}}(l(Y, z)|X = x) \tag{8.40}$$

Theorem

For all $f : \mathbb{R}^p \to \mathcal{Y}$,

$$R(f) \geq r^* \tag{8.41}$$

with $R(f) - R^*$ called the excess risk.

Example - Regression

For the squared loss $l(a, b) := (a - b)^2$, and $\mathcal{Y} = \mathbb{R}$, the Bayes predictor (=regression function) and Bayes risk are

$$f^*(x) = \mathbb{E}_{Y|X}[Y|X = x] \qquad R^* = \mathbb{E}_X\mathbb{E}_{Y|X}\left[(Y - \mathbb{E}_{Y|X}[Y|X])^2|X\right] \tag{8.42}$$

The optimal prediction of Y on x is thus the mean of Y conditioned to $X = x$. It is called a conditional mean.

Example - Classification

For the 0/1 loss $l(a, b) := I(a \neq b)$, and $\mathcal{Y} = \mathcal{G} = \{g_1, \ldots, g_k\}$, the Bayes predictor (= Bayes classifier) and Bayes risk are

$$f^*(x) = \arg\max_{z\in\mathcal{G}} p(Y = z|X = x) \qquad R^*1 - \mathbb{E}_X \max_{z\in\mathcal{G}} p(Y = z|X) \tag{8.43}$$

The optimal classifier thus selects the most likely class given x.

k-NN and Bayes

The Bayes classifier is optimal, by definition, for classifications with

$$f^*(x) = \arg\max_{z \in \mathcal{G}} p(Y = z | X = x) \tag{8.44}$$

and the $k-$NN solution approaches f^* for N, k large and k/N small. However, the impact of the feature space dimension p is not forgotten.

8.9 Model selection and Bias-Variance tradeoff

8.9.1 Bias-Variance decomposition

Let us consider the following special case:

- A dataset $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^N$ sampled from $p(X, Y)$ with

$$Y(X) = f(X) + \varepsilon \qquad \mathbb{E}(\varepsilon) = 0 \qquad \mathbb{V}(\varepsilon) = \sigma_\varepsilon^2 \tag{8.45}$$

for a certain noise ε and a deterministic, unknown function f. i.e., given x_i, $y_i = f(x_i) + \varepsilon_i$, with $\varepsilon_i \sim_{i.d.d.} \varepsilon$.

- An estimate $\hat{f}(\cdot) = f_{\hat{\beta}}(\cdot)$ of f, with $\hat{\beta}$ the hyperparameters trained from $\mathcal{T}$.

We analyze the expected prediction error $EPE(x_0)$ of $\hat{f}$ on $x_0 \notin \{x_i\}_{i=1}^N$:

$$EPE(x_0) \coloneqq \mathbb{E}\left[(Y(x_0) - \hat{f}(x_0))^2 | X = x_0\right] \tag{8.46}$$

with $\mathbb{E}$ considered on both $\mathcal{T}$ and Y. We call the term $Y(x_0) - \hat{f}(x_0)$ the variability of f.
We know that, conditionally to $X = x_0$, $Y(x_0) = f(x_0) + \varepsilon_0$, and

$$Y(x_0) - \hat{f}(x_0) = \varepsilon_0 + (f(x_0) - \mathbb{E}(\hat{f}(x_0)) + (\mathbb{E}\hat{f}(x_0) - \hat{f}(x_0)) \tag{8.47}$$

Therefore, since, the first and last term are independent with null mean,

$$EPE(x_0) = \mathbb{E}\left[(Y(x_0) - \hat{f}(x_0))^2 | X = x_0\right] = \sigma_\varepsilon^2 + \left(f(x_0) - \mathbb{E}\hat{f}(x_0)\right)^2 + \mathbb{E}\left(\hat{f}(x_0) - \mathbb{E}\hat{f}(x_0)\right)^2 \tag{8.48}$$

The first term is called the noise variance, the second is the square of the bias of $\hat{f}(x_0)$, and the last is the variance $\mathbb{V}_{\mathcal{T}}(\hat{f}(x_0))$.

- Irreducible term : σ_ε^2 is the variance of the output of the new test point x_0.

- Reducible term : The variance is the variability with respect to the dataset.

- Reducible term : The bias is the error of the model and measures how good it can be at predicting f.

→ N.B.: the bias decreases and the variance increases when the model flexibility (N/k) increases.

Bias-Variance tradeoff for k-NN

Let us consider that

- the positions $\{x_i\}_{i=1}^N \subset \mathbb{R}^p$ are fixed, and $x_0 \notin \{x_i\}_{i=1}^N$;
- $\mathcal{T} = \{(x_i, Y(x_i) = f(x_i) + \varepsilon_i)\}_{i=1}^N$, with $\varepsilon_i \sim_{i.d.d.} N(0, \sigma_\varepsilon^2)$;
- $N_k(x_0) = \{x_{(l)}\}_{l=1}^k$, the fixed k neighbors of x_0.

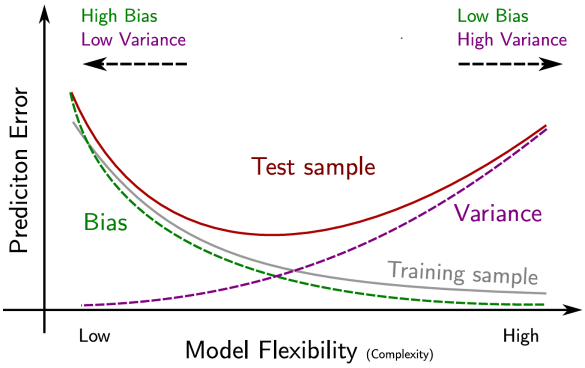
1. Regarding the bias, for $\{x_{(l)}\}_{l=1}^k$,

$$\left(\text{Bias}\hat{f}(x_0)\right)^2 := \left(f(x_0) - \mathbb{E}\hat{f}(x_0)\right)^2 = \left(f(x_0) - \frac{1}{k}\sum_{l=1}^k f(x_{(l)})\right)^2 \tag{8.49}$$

2. Regarding the variance,

$$\mathbb{V}_{\mathcal{T}}(\hat{f}(x_0)) = \mathbb{V}\left(\frac{1}{k}\sum_{l=1}^k Y(x_{(l)})\right) = \frac{\sigma_\varepsilon^2}{k} \tag{8.50}$$

Schematically, as the test sample has a U-shape as it is the sum of two curves:



Data science - Unsupervised Learning

9.1 Introduction

The difference between supervised learning and unsupervised learning is that SL learns buy examples, while UL learns by observations. The goal of UL is not to do prediction, but rather to discover arrangements, clusters, patterns,etc.

9.2 Principal Component Analysis (PCA)

9.2.1 Dimensionality Reduction Problem

We have p features : $X = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$, sampled on N observations in $\mathcal{X} := \{x_i \in \mathbb{R}^p\}_{i=1}^N$. We assume that $\sum_{i=1}^N x_i = 0$ for simplicity (centering the data).

The dimensionality reduction problem is transforming $\mathcal{X} \subset \mathbb{R}^p$ into $\mathcal{X}' \subset \mathbb{R}^{p'}$, with $p' \ll p$, where $\mathcal{X}'$ preserves essential information about each element of $\mathcal{X}$.
This problem is solved by the PCA: it is an unsupervised algorithm to find the best axes to represent a dataset $\mathcal{X}$, i.e. find the axes that maximize the variance of the dataset $\mathcal{X}$.

9.2.2 Directional variance

Given a direction $\varphi := (\varphi_1, \ldots, \varphi_p)^T \in \mathbb{R}^p$, with $\|\varphi\|^2 := \sum_i \varphi_i^2 = 1$. We compute the projection of all points of $\mathcal{X}$ onto φ:

$$\mathcal{Z}(\varphi) = \{z_i := \varphi^T x_i\}_{i=1}^N \subset \mathbb{R} \tag{9.1}$$

and the variance V of all projections:

$$V(\varphi) := \frac{1}{N} \sum_{i=1}^N z_i^2 = \frac{1}{N} \varphi^T X^T X \varphi \tag{9.2}$$

Therefore, the first principal component of $\mathcal{X}$ is

$$\varphi^{(1)} = \arg \max_{\varphi : \|\varphi\| = 1} \frac{1}{N} \varphi^T X^T X \varphi \tag{9.3}$$

Finding the second,. . . , k-th principal component works the same way, with the additional constraint that the i-th component must be orthogonal to all the previous components, so that they are uncorrelated.

In summary, if we compute p principal components, we get an orthonormal basis of $\mathbb{R}^p$. This is an alternative representation of the canonical basis and it is adapted to the structure of $\mathcal{X}$.

9.2.3 Interpretation

Given $1 \leq k \leq \min(p, N-1)$,

- The k first PCs of $\mathcal{X}$ are the k axes $\varphi^{(j)} \subset \mathbb{R}^p$.
- The k-dimensional space with highest variance S_k is the vector subspace of the φ_j.
- S_k is the closest to the observations in $\mathcal{X}$.

9.2.4 Inverse transform

The forward transform is, given an observation x_i:

$$x_i \in \mathbb{R}^p \to z_i = (z_{i,1} = \varphi_1^T x, \dots, z_{i,k} = \varphi_k^T x_i)^T \in \mathbb{R}^k \tag{9.4}$$

Since the $\varphi^{(j)}$ form an orthonormal basis, the scores can be pushed back to $\mathbb{R}^p$ with

$$x_i \in \mathbb{R}^p \to x_i' := \sum_{j=1}^{k} z_{i,j} \cdot \phi_j \in \mathbb{R}^p \tag{9.5}$$

→ N.B.: doing the forward, then inverse transform suppresses a lot of noise.

To determine how many PCs to keep, we compute the proportion of the variance explained (PVE). Assuming centered data,

- For fixed $j \in [p]$ and $i \in [N]$, $x_{ij} \to j$-th features of the i-th sample of $X \in \mathbb{R}^p$.
- The dataset of all j-th features $\to X_j := \{x_{ij}\}_{i=1}^N \subset \mathbb{R}$.
- The total variability of $\mathcal{X}$ is

$$V_{\mathcal{X}} := \sum_{j=1}^{p} V_{\mathcal{X}_j} = \sum_{j=1}^{p} \frac{1}{N} \sum_{i=1}^{N} x_{ij}^2 \tag{9.6}$$

We also define the PVE of every component:

$$PVE(j) := \frac{V(\varphi^{(j)})}{V_{\mathcal{X}}} \leq 1 \tag{9.7}$$

Therefore, the cumulative PVE (cPVE) as a function of j is

$$cPVE(j) := \sum_{l=1}^{j} PVE(l) \leq 1 \tag{9.8}$$

We choose the value of the parameter k depending on the cPVE accuracy we want.

It is important that the variables have a zero mean and are scaled before computing the PCA.

9.3 Clustering methods

9.3.1 Clustering with K-Means

Given an unlabeled dataset $\mathcal{X} = \{x_i\}_{i=1}^N \subset \mathbb{R}^p$. The K-Means aims at partitioning $\mathcal{X}$ into K distinct, non-overlapping clusters.

We define a set of K clusters $\mathcal{C} := \{\mathcal{C}_1, \dots, \mathcal{C}_k\}$ such that

- $\mathcal{C}_j \subset [N]$, with $\mathcal{C}_j$ the indices of the points in the j-th cluster.
- $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$ if $i \neq j$ and $\cup_j \mathcal{C}_j = [N]$.

For the K-Means method, a good clustering $\mathcal{C}$ minimizes the in-cluster variations $V_{in}(\mathcal{C}) := \sum_{j=1}^K V_c(\mathcal{C}_j$, with $V_c(\mathcal{C}_j) := \frac{1}{|\mathcal{C}_j|} \sum_{i,i' \in \mathcal{C}_j} \|x_i - x_{i'}\|^2$.

Therefore, K-Means aims at minimizing the total variation of $\mathcal{C}$, i.e.

$$\hat{C} = \arg\min_{\mathcal{C}} \sum_{j=1}^K V_c(\mathcal{C}_j) = \arg\min_{\mathcal{C}} \sum_{j=1}^K \frac{1}{|\mathcal{C}_j|} \sum_{i,i' \in \mathcal{C}_j} \|x_i - x_{i'}\|^2 \tag{9.9}$$

→ N.B.: we have $\sum_{j=1}^K \frac{1}{|\mathcal{C}_j|} \sum_{i,i' \in \mathcal{C}_j} \|x_i - x_{i'}\|^2 = 2 \sum_{j=1}^K \sum_{i \in \mathcal{C}_|} \|x_i - c_j\|^2$, with the centroid $c_j := \frac{1}{|\mathcal{C}_j|} \sum_{i \in \mathcal{C}_j} x_i$ of the j-th cluster.

Therefore, K-Means minimizes the distances to the centroids $\{c_j\}_{j=1}^K$.

Lloyd-Max algorithm

Given the number of clusters $K \in \mathbb{N}$,

- initialization: randomly assign K centroids from $\mathcal{C}$.
- Update $\mathcal{C}$: for all $1 \leq j \leq K$,
  - $\mathcal{C}_j \leftarrow \{i \in [N] : \|x_i - c_j\| \leq \|x_i - c_{j'}\| \qquad \forall j' \in [K]\}$.
  - This means that $\mathcal{C}_j$ contains the indices of the points that are closer to c_j than to any other centroid.
- Update centroids: $c_j \leftarrow \frac{1}{|\mathcal{C}_j|} \sum_{i \in \mathcal{C}_j} x_i$.
- Iterate until convergence.

For this algorithm, we need to know in advance the number of clusters that are needed.

Selecting the number of clusters

A possible solution is the silhouette score analysis:
Given $i^* \in \mathcal{C}_{j^*} \in \mathcal{C}$, we define:

- The mean intra-cluster distance:

$$a(x_{i^*}) = \frac{1}{|\mathcal{C}_{j^*}| - 1} \sum_{i \neq i^* i \in \mathcal{C}_{j^*}} d(x_{i^*}, x_i) \tag{9.10}$$

- The mean nearest-cluster distance:

$$b(x_{i^*}) = \min_{j \neq j^*} \frac{1}{|\mathcal{C}_j|} \sum_{i \in \mathcal{C}_j} d(x_{i^*}, x_i) \tag{9.11}$$

The silhouette score of x_{i^*} is

$$s(x_i^*) = \frac{b(x_{i^*} - a(x_{i^*}}{\max(a(x_{i^*}), b(x_{i^*}))} \in [1-, 1] \tag{9.12}$$