Aprenentatge Automàtic 1

GCED

Lluís A. Belanche

belanche@cs.upc.edu

Soft Computing Research Group
Dept. de Ciències de la Computació (Computer Science)

Universitat Politècnica de Catalunya

2019-2020

LECTURE 6: Classification theory and linear classification models (III)

Discriminative classifiers

Outline

1. **Maximum Likelihood** (ML) framework (I)
   - Gentle motivation
   - Example 1: ML method for the Gaussian
   - Example 2: ML method for coin flipping
   - Example 3: ML method for a biased die
2. **Generalized Linear Models (GLM)**
   - Generative vs. discriminative classifiers
   - Basic GLM setting
   - The Logistic Regression model
   - The Poisson regression model
3. **Maximum Likelihood** (ML) framework (II)
   - Maximum Likelihood theory
   - Bias and Variance of a ML estimate
   - Some properties of ML
   - Examples

Maximum likelihood

Gentle motivation

We have a r.v. X (e.g., the height of a randomly chosen Dutch)

The population has some distribution, which is assumed to have a special form. A common choice for a continuous distribution is the Gaussian (or normal) density*:

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

and written $X \sim N(x;\mu,\sigma^2)$, where μ, the mean, and σ^2, the variance, are the **parameters** of the distribution

---

(*) The probability that $X \in (x - \Delta x, x + \Delta x)$ approaches $2f(x)\Delta x$ as $\Delta x \longrightarrow 0$.

Maximum likelihood

Gentle motivation

- Suppose we take an i.i.d sample $D = \{x_1, \dots, x_N\}$ of the r.v. X
- From the sample, we wish to **estimate** μ (it could be σ^2)
- It is not clear *a priori* what is the best way to do this:

1. the average of D?
2. the median of D?
3. the average of the minimum and the maximum in D?

Maximum likelihood

Example 1: the Gaussian

The **likelihood** of observing a particular x_n is $f(x_n; \mu, \sigma^2)$

The **likelihood** of seeing all the sample D is $\prod_{n=1}^{N} f(x_n; \mu, \sigma^2)$

Viewing this as *a function of the parameters*, we define

$$\mathcal{L}\left(\mu, \sigma^2\right) = P(D \mid \mu, \sigma^2) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x_n - \mu)^2}{\sigma^2}\right)$$

"how likely it is that the population has parameters μ and σ^2 *given* the observed data sample D"

The **maximum likelihood** estimates for the parameters are the values $\widehat{\mu}$ and $\widehat{\sigma}^2$ that maximize $\mathcal{L}\left(\mu, \sigma^2; D\right)$

Maximum likelihood

Example 1: the Gaussian

It is sometimes convenient (and equivalent) to maximize the "log-likelihood":

$$l(\mu, \sigma^2) := \ln \mathcal{L}\left(\mu, \sigma^2\right) = \sum_{n=1}^{N} \ln f(x_n; \mu, \sigma^2)$$

In the Gaussian case, we get:

$$l(\mu, \sigma^2) = \sum_{n=1}^{N} \left[\ln \frac{1}{\sqrt{2\pi}} - \ln \sigma - \frac{1}{2} \left(\frac{x_n - \mu}{\sigma} \right)^2 \right]$$

Maximum likelihood

Example 1: the Gaussian

1. We compute $\dfrac{\partial l}{\partial \mu} = \dfrac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu)$,

2. We make $\dfrac{\partial l}{\partial \mu} = 0$, obtaining $\widehat{\mu} = \dfrac{1}{N} \sum_{n=1}^{N} x_n$, the average of the sample

3. Also $\dfrac{\partial^2 l}{\partial \mu \partial \mu} = -\dfrac{N}{\sigma^2} < 0$, and therefore we have found a maximum

---

The estimate for the variance is $\widehat{\sigma}^2 = \dfrac{1}{N} \sum_{n=1}^{N} (x_n - \widehat{\mu})^2$

Maximum likelihood

Example 2: coin flipping

- Suppose we flip a coin that turns up heads with probability p. Find the ML estimate for p

- We take a sample $D = \{x_1, \ldots, x_N\}$ out of N flips and get n_1 heads and $N - n_1$ tails. The number of heads follows a binomial $B(N, p)$

- The likelihood is $\mathcal{L}(p) = \binom{N}{n_1} p^{n_1} (1-p)^{N-n_1}$

- The log-likelihood is $l(p) = \ln \binom{N}{n_1} + n_1 \ln p + (N - n_1) \ln(1-p)$

$$\frac{\partial l}{\partial p} = \frac{n_1}{p} - \frac{N - n_1}{1 - p} = 0, \qquad \text{therefore} \qquad \widehat{p} = \frac{n_1}{N}$$

Maximum likelihood

Example 3: biased die

- Estimate the probabilities $\pi_1, \ldots, \pi_6$ of a (possibly biased) die rolling numbers $1, \ldots, 6$

- We observe n_1 ones, n_2 twos, etc, in N rolls

- The ML estimates we get are $\widehat{\pi}_k = \dfrac{n_k}{N}$

  (to solve, one must add the constraint $\displaystyle\sum_{k=1}^{6} \pi_k = 1$)

Generalized Linear Models (GLM)

- Very general and classical technique for fitting **linear models**

- Genuine representatives of **discriminative** methods

- Work for many **target types**:
  - binary (two-class) and nominal (multi-class)
  - proportions and counts
  - ordinal (ordered classes)
  - continuous

- Admit **general predictors** (categorical ones are binarized)

Generalized Linear Models

- A GLM is a linear predictor of a convenient function of the **conditional expectation** of the target variable:

$$h(\mathbb{E}[T_n|\boldsymbol{X_n}]) = \boldsymbol{\beta}^\top \boldsymbol{X}_n + \beta_0$$

- This convenient function h is typically smooth and invertible and called the **link function**
- We **optimize** the $\boldsymbol{\beta}$ and β_0 parameters directly (without distributional assumptions on the $\boldsymbol{x}$)
- The T_n are taken as independent and drawn from a distribution of the **exponential family** (Poisson, Gaussian, Bernoulli, Gamma, ...)

Generalized Linear Models

The modeller chooses a suitable distribution for T_n given $\boldsymbol{X_n}$:

1. Gaussian distribution (h is identity): **linear regression**

2. Bernoulli distribution (h is logit): **logistic regression**

3. Poisson distribution (h is ln): **Poisson regression**

- This generality comes at a cost: in general we need an iterative procedure for the optimization of the $\boldsymbol{\beta}$ and β_0 parameters

- A popular procedure is to set it up as a ML problem and use a preferred numerical optimization method (e.g. Newton-Raphson)

Logistic regression

Introduction

We are in the case of binary classification ($K = 2$). We **model** the posterior probability for class ω_1 as:

$$P(\omega_1|\boldsymbol{x}) = g(\boldsymbol{\beta}^\top \boldsymbol{x} + \beta_0)$$

where $g(z) = \dfrac{\exp(z)}{1 + \exp(z)} = \dfrac{1}{1 + \exp(-z)}$ is the **logistic function**

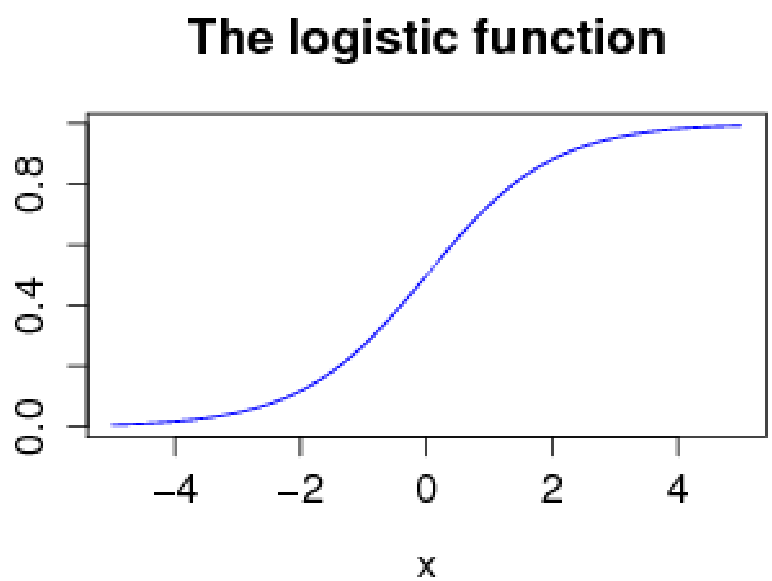
---

obviously $P(\omega_2|\boldsymbol{x}) = 1 - P(\omega_1|\boldsymbol{x}) = 1 - g(\boldsymbol{\beta}^\top \boldsymbol{x} + \beta_0)$

Logistic regression

The logistic function

The logistic function is a C^∞ function $g : \mathbb{R} \longrightarrow (0,1)$



This function is a bijection (one-to-one), with inverse $g^{-1}(z) = \ln\left(\dfrac{z}{1-z}\right)$, for $z \in (0,1)$ (the **logit function**)

Logistic regression

Logistic Regression as a GLM

Each $T_n \sim \text{Ber}(p_n)$, where $p_n = g(\boldsymbol{\beta}^\top \boldsymbol{X}_n + \beta_0)$,

$$\Rightarrow P(T_n|\boldsymbol{X}_n, \boldsymbol{\beta}) = \begin{cases} p_n & \text{if } T_n = 1 \ (\boldsymbol{X}_n \in \omega_1) \\ 1 - p_n & \text{if } T_n = 0 \ (\boldsymbol{X}_n \in \omega_2) \end{cases}$$

$$= p_n^{T_n}(1 - p_n)^{(1-T_n)}$$

Note $g(\boldsymbol{\beta}^\top \boldsymbol{X}_n + \beta_0) = \mathbb{E}[T_n] = p_n$ (GLM setting)

$\rightarrow$ we are identifying p_n with $P(\omega_1|\boldsymbol{X}_n)$

Logistic regression

Interpretation of our model

"The log of the odds is a linear function of the predictors"

Since $P(\omega_1|\boldsymbol{X}) = g(\boldsymbol{\beta}^\top \boldsymbol{X} + \beta_0)$

we have

$$\ln\left(\frac{P(\omega_1|\boldsymbol{X})}{P(\omega_2|\boldsymbol{X})}\right) = \ln\left(\frac{P(\omega_1|\boldsymbol{X})}{1 - P(\omega_1|\boldsymbol{X})}\right) = \text{logit}(P(\omega_1|\boldsymbol{X})) = \boldsymbol{\beta}^\top \boldsymbol{X} + \beta_0$$

Logistic regression

Let's go for the parameters

Suppose we have an i.i.d. sample of N *labelled* observations $S = \{(\boldsymbol{x}_n, t_n)\}_{n=1,\ldots,N}$, where $\boldsymbol{x}_n \in \mathbb{R}^d, t_n \in \{0, 1\}$

1. The first thing we note is that we have $d + 1$ parameters to fit

2. In the "equivalent" generative case (LDA), we had $\frac{d(d+1)}{2} + 2d$

Let's re-write $P(\omega_1|\boldsymbol{X}) = g(\boldsymbol{\beta}^\top \boldsymbol{X})$

with $\boldsymbol{X} = (1, X_1, \ldots, X_d)^\top$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^\top$.

Logistic regression

ML for the Logistic Regression

$$
\begin{aligned}
l(\boldsymbol{\beta}) &= \ln \mathcal{L}(\boldsymbol{\beta}) \\
&= \ln P(\{t_n\}_{n=1}^{N} | \{\boldsymbol{x}_n\}_{n=1}^{N}, \boldsymbol{\beta}) \\
&= \ln \prod_{n=1}^{N} P(t_n | \boldsymbol{x}_n, \boldsymbol{\beta}) \\
&= \sum_{n=1}^{N} \ln P(t_n | \boldsymbol{x}_n, \boldsymbol{\beta}) \\
&= \sum_{n=1}^{N} \ln \left(g(\boldsymbol{\beta}^\top \boldsymbol{x}_n)^{t_n} \Big(1 - g(\boldsymbol{\beta}^\top \boldsymbol{x}_n)\Big)^{(1-t_n)} \right) \\
&= \sum_{n=1}^{N} \ln \left((p_n)^{t_n} (1 - p_n)^{(1-t_n)} \right), \qquad p_n = g(\boldsymbol{\beta}^\top \boldsymbol{x}_n)
\end{aligned}
$$

Logistic regression

ML for the Logistic Regression

Now

$$
\begin{aligned}
(p_n)^{t_n}(1-p_n)^{(1-t_n)} &= \left(\frac{p_n}{1-p_n}\right)^{t_n}(1-p_n) \\
&= \Big(\exp(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}_n)\Big)^{t_n}\Big(1+\exp(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}_n)\Big)^{-1}
\end{aligned}
$$

Therefore

$$
l(\boldsymbol{\beta}) = \sum_{n=1}^{N}\left[t_n\boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}_n - \ln\Big(1+\exp(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}_n)\Big)\right]
$$

Logistic regression

Newton-Raphson

The ML of the logistic regression does not have a closed-form solution:

$$\begin{aligned}
\boldsymbol{\beta}^{k+1} &= \boldsymbol{\beta}^k - \left(\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\right)^{-T}\left(\frac{\partial l}{\partial\boldsymbol{\beta}}\right) \\
&= \boldsymbol{\beta}^k + (X^\top W X)^{-1} X^\top(\boldsymbol{t}-\boldsymbol{p}) \\
&= (X^\top W X)^{-1} X^\top W \boldsymbol{z}, \qquad \boldsymbol{z} = X\boldsymbol{\beta}^\top + W^{-1}(\boldsymbol{t}-\boldsymbol{p})
\end{aligned}$$

where:

X is the matrix of the $\{\boldsymbol{x}_n\}$

$W = \mathrm{diag}(p_n(1-p_n)),\ n = 1,\ldots,N$

$\boldsymbol{t} = (t_1,\ldots,t_N)^\top,\ \boldsymbol{p} = (p_1,\ldots,p_N)^\top$

since:

$$\frac{\partial l}{\partial\boldsymbol{\beta}} = X^\top(\boldsymbol{t}-\boldsymbol{p})$$

$$\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} = -X^\top W X$$

Logistic regression

Iterated Reweighted Least Squares (IRLS)

1. Initialize $\beta_0 \leftarrow \ln\left(\dfrac{P(\omega_1)}{1 - P(\omega_1)}\right)$ and $\beta_i \leftarrow 0, i = 1, \ldots, d$ (*null model*)

2. Iterate until convergence:

   a) Update $\boldsymbol{p} \leftarrow (p_1, \ldots, p_N)^\top$, where $p_n = g(\boldsymbol{\beta}^\top \boldsymbol{x}_n)$

   b) Update $W^{-1} \leftarrow \text{diag}\left(\frac{1}{p_n(1-p_n)}\right),\ n = 1, \ldots, N$

   c) Update $\boldsymbol{z} \leftarrow X\boldsymbol{\beta}^\top + W^{-1}(\boldsymbol{t} - \boldsymbol{p})$

   d) Update $\boldsymbol{\beta} \leftarrow (X^\top W X)^{-1} X^\top W \boldsymbol{z}$

3. return $\widehat{\boldsymbol{\beta}}$

Logistic regression

The Deviance and the AIC

In the context of Generalized Linear Models,

$$-2l(\widehat{\boldsymbol{\beta}}) = -2\ln\mathcal{L}(\widehat{\boldsymbol{\beta}}) \sim \chi^2_{v=N-d-1}$$

is called the **deviance** (in ML, this is the **error**)

Null deviance: deviance of the null model (just with constant term)
Residual deviance: deviance of the proposed model
AIC: deviance with complexity penalization $-2l(\widehat{\boldsymbol{\beta}}) + 2d$

Actually, $2d \approx 2\|\widehat{\boldsymbol{\beta}}\|_0$ is a rudimentary form of **regularization**

An example of R's glm() in action

```
glm (formula = chd ~ age + sbp + ldl + adiposity + alcohol + tobacco +
    obesity + famhist + typea, family = binomial, data = SAheart.learn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9287  -0.8283  -0.3762   0.8983   2.4722

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -7.066884   1.654749  -4.271 1.95e-05 ***
age             0.060443   0.015789   3.828 0.000129 ***
sbp             0.005366   0.007127   0.753 0.451531
ldl             0.195473   0.078493   2.490 0.012763 *
adiposity      -0.009770   0.035926  -0.272 0.785658
alcohol        -0.001653   0.005979  -0.277 0.782146
tobacco         0.090333   0.033137   2.726 0.006411 **
obesity        -0.028118   0.053949  -0.521 0.602229
famhistPresent  0.912713   0.283802   3.216 0.001300 **
typea           0.041579   0.015356   2.708 0.006777 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null deviance: 401.48  Residual deviance: 308.55
AIC: 328.55            Number of Fisher Scoring iterations: 5
```

Logistic regression

An example of R's `glm()` in action

Use of the AIC to simplify the model:

```
Coefficients:
   (Intercept)            age            ldl        tobacco  famhistPresent
      -7.12058        0.05943        0.17677        0.08927         0.89552
         typea
       0.04012

Null Deviance:     401.5
Residual Deviance:  310.1  AIC: 322.1
```

Logistic regression

Interpreting the coefficients

$$\mathsf{LOGODDS}(\boldsymbol{x}_0) = \ln\left(\frac{P(\omega_1|\boldsymbol{x}_0)}{P(\omega_2|\boldsymbol{x}_0)}\right) = \boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}_0 + \beta_0$$

$$\mathsf{ODDS}(\boldsymbol{x}_0) = \frac{P(\omega_1|\boldsymbol{x}_0)}{P(\omega_2|\boldsymbol{x}_0)} = \exp(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}_0 + \beta_0)$$

Define $\mathbf{1}_i = (0, \ldots, \overset{i)}{1}, \ldots, 0)^\mathsf{T}$, so $\boldsymbol{x}_0 + \mathbf{1}_i = (x_{01}, \ldots, x_{0i} + 1, \ldots, x_{0N})^\mathsf{T}$

$$\frac{\mathsf{ODDS}(\boldsymbol{x}_0 + \mathbf{1}_i)}{\mathsf{ODDS}(\boldsymbol{x}_0)} = \exp\Big((\boldsymbol{\beta}^\mathsf{T}(\boldsymbol{x}_0 + \mathbf{1}_i - \boldsymbol{x}_0)\Big) = \exp(\beta_i)$$

Poisson regression

Introduction

In many statistical studies, one tries to relate a count to some scientific variables:

1. Number of cardio-vascular accidents among people over 60 in a US state $\sim$ average income in the state

2. Number of bicycles in a Danish household $\sim$ distance to the city centre

3. Number of incoming calls to a complaints number $\sim$ hourly interval

When the response is a count which does not have any natural upper bound, the logistic regression is not appropriate.

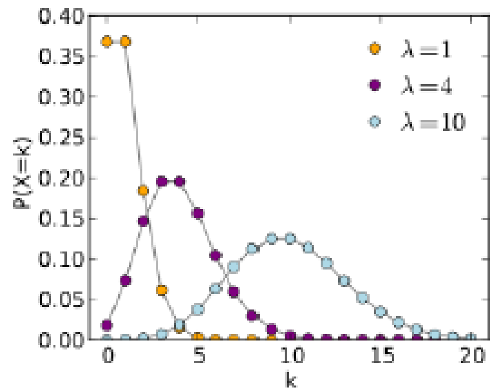
$\rightarrow$ The **Poisson regression** is a natural alternative

Poisson regression

The Poisson distribution

This is a discrete distribution $X \sim \text{Pois}(\lambda)$ with probability mass function:


Simeon Poisson 1781-1840



$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad \lambda > 0 \in \mathbb{R}, k = 0, 1, 2, \ldots$$

Poisson regression

- We consider independent Poisson random variables $T_1, \ldots, T_N$ with $T_n \sim \text{Pois}(\lambda_n)$. We know that $\mathbb{E}[T_n] = \lambda_n$

- We have an i.i.d. sample of N observations $\boldsymbol{x}_n \in \mathbb{R}^d$ (*e.g.* distance to the city centre)

- We have a corresponding sample $t_1, \ldots, t_N$, where each t_n is drawn from T_n (*e.g.* number of bicycles)

- The idea is to model λ_n as $\exp(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}_n + \beta_0)$ (the link function is ln)

- The Poisson regression model is $T_n \sim \text{Pois}\Big(\exp(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}_n + \beta_0)\Big)$ or $\ln \lambda_n = \boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}_n + \beta_0$

Poisson regression

ML for the Poisson regression

- Proceeding like the previous case for Logistic regression, we arrive at:

$$l(\boldsymbol{\beta}) = \sum_{n=1}^{N} \Big[-\mathsf{exp}(\beta^{\mathsf{T}}\boldsymbol{x}_n) + t_n\beta^{\mathsf{T}}\boldsymbol{x}_n - \mathsf{ln}(t_n!) \Big]$$

- Again, this expression has no closed-form solution; however, we can still use Newton-Raphson (because $-l$ is convex)

Maximum likelihood

Some theory

1. Consider a random vector $X_1, \cdots, X_N$ each of which having mass (or density) $f(x_n; \theta)$ and a random sample $\{x_1, \cdots, x_N\}$ thereof

2. The **likelihood function** is defined as the product:

   $$\mathcal{L}(\theta) = f(x_1;\theta)f(x_2;\theta)\cdots f(x_n;\theta) = \prod_{n=1}^{N} f(x_n;\theta)$$

   (**note:** x_i are numbers while X_i are random variables)

3. The likelihood is considered a function of θ for fixed data (whereas the mass or density are considered a function of x for fixed θ)

   It can be read as the probability of observing the available sample for different values of the θ parameter

Maximum likelihood

Some theory

- The **ML estimator** (MLE) $\widehat{\Theta}(X_1, \cdots, X_N)$ is the *function* of the random vector that maximizes the likelihood with respect to θ

- The **ML estimate** (MLe) $\widehat{\theta}_N = \widehat{\Theta}(x_1, \cdots, x_N)$ gives a *point* estimation for the available data

  (the MLE is a function while the MLe is a number)

- How good are these estimates?

  - **Unbiasedness**: is the expected value of the estimate the number being estimated?

  - **Efficiency**: is the variance of some estimate smaller than that of another estimate?

Maximum likelihood

Bias

Let $\widehat{\theta}_N = \widehat{\Theta}(x_1, \cdots, x_N)$. The **bias** of $\widehat{\theta}_N$ is:

$$\text{Bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta}_N] - \theta$$

We say that an estimate $\widehat{\theta}_N$ is **unbiased** if:

$$\mathbb{E}[\widehat{\theta}_N] = \theta$$

the expected value is taken over all possible samples of a given size N

Maximum likelihood

Variance

The **variance** of $\widehat{\theta}_N$ is:

$$\mathsf{Var}(\widehat{\theta}_N) = \mathbb{E}[\ (\widehat{\theta}_N - \mathbb{E}[\widehat{\theta}_N])^2\ ]$$

We say that an estimate $\widehat{\theta}$ is **more efficient** than an estimate $\widehat{\theta}'$ if $\mathsf{Var}(\widehat{\theta}) < \mathsf{Var}(\widehat{\theta}')$

There is a lower limit on the variance of estimates (Cramér-Rao bound)

Maximum likelihood

Consistency

We say that an estimate $\widehat{\theta}_N$ is **consistent** if it tends in probability to θ as $N \longrightarrow \infty$. Formally:

$$\forall \epsilon > 0, \ \lim_{N \to \infty} P(|\widehat{\theta}_N - \theta| < \epsilon) = 1$$

Consistency implies that:

1. Bias decreases with sample size

2. Variance decreases with sample size

Maximum likelihood

Some general properties of ML

Assymptotic unbiasedness :

- Any ML estimate is at least asymptotically unbiased (hence, when the sample size is large, it is approximately unbiased)
- It is the best estimate as the sample size N becomes large (other estimates can be better for small N)

Asymptotic efficiency : The ML estimate is the more efficient among those that are consistent

Invariance : If $\hat{\Theta}$ is the MLE of θ and g is bijective (one-to-one), then the MLE of $g(\Theta)$ is $g(\hat{\Theta})$

Maximum likelihood

Example 1: the Gaussian

We had previously found that:

- The ML estimate for the mean μ of a Gaussian is $\widehat{\mu}_N = \frac{1}{N}\sum_{n=1}^{N} x_n$, the sample mean

- The ML estimate for the variance σ^2 of a Gaussian is $\widehat{\sigma}_N^2 = \frac{1}{N}\sum_{n=1}^{N} (x_n - \widehat{\mu}_N)^2$, the sample variance

How good are these estimates?

Do they have a bias? What is their variance?
Are they consistent? Efficient?

Maximum likelihood

Example: the Gaussian

$$
\begin{aligned}
\text{Bias}(\widehat{\mu}_N) &= \mathbb{E}[\widehat{\mu}_N] - \mu \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} x_n\right] - \mu \\
&= \left(\frac{1}{N}\sum_{n=1}^{N} \mathbb{E}[x_n]\right) - \mu \\
&= \left(\frac{1}{N}\sum_{n=1}^{N} \mu\right) - \mu \\
&= \frac{N\mu}{N} - \mu = 0
\end{aligned}
$$

Maximum likelihood

Example: the Gaussian

$$\mathrm{Var}(\widehat{\mu}_N) = \mathbb{E}\Big[(\widehat{\mu}_N - \mathbb{E}[\widehat{\mu}_N])^2\Big]$$

$$\begin{aligned}
\mathrm{Var}(\widehat{\mu}_N) &= \mathrm{Var}\left[\frac{1}{N}\sum_{n=1}^{N} x_n\right] \\
&= \frac{1}{N^2}\sum_{n=1}^{N} \mathrm{Var}[x_n] \\
&= \frac{1}{N^2}\sum_{n=1}^{N} \sigma^2 = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}
\end{aligned}$$

Since $\displaystyle \lim_{N\to\infty} \frac{\sigma^2}{N} = 0$, it is **unbiased** and **consistent**

Maximum likelihood

Example: the Gaussian

Since

$$\mathbb{E}[\widehat{\sigma}_N^2] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \widehat{\mu}_N)^2\right] = \left(\frac{N-1}{N}\right)\sigma^2$$

this estimator is biased! (it underestimates the true value by a factor of $\frac{N-1}{N}$)

Consider $\tilde{\sigma}_N^2 = \frac{N}{N-1}\widehat{\sigma}_N^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \widehat{\mu}_N)^2$

Now $\mathbb{E}[\tilde{\sigma}_N^2] = \mathbb{E}\left[\frac{N}{N-1}\widehat{\sigma}_N^2\right] = \frac{N}{N-1}\mathbb{E}[\widehat{\sigma}_N^2] = \frac{N}{N-1}\left(\frac{N-1}{N}\right)\sigma^2 = \sigma^2$

The expression for Var$[\widehat{\sigma}_N^2]$ (or Var$[\tilde{\sigma}_N^2]$) is more complex but is of the form $O(\frac{1}{N})$. Therefore both $\widehat{\sigma}_N^2$ and $\tilde{\sigma}_N^2$ are **consistent** estimators of σ^2.

Maximum likelihood

Example: the Gaussian

It is important to note that unbiased and consistent estimators are not always the best estimators! Sometimes efficiency is a must!

When μ is unknown, the estimator:

$$\bar{\sigma}_N^2 = k_N^2 \cdot \tilde{\sigma}_N^2 \approx \frac{1}{N - 1{,}45} \sum_{n=1}^{N} (x_n - \widehat{\mu}_N)^2$$

–with $k_N = \sqrt{\frac{N-1}{2}}\Gamma(\frac{N-1}{2})/\Gamma(\frac{N}{2})$, being Γ the gamma function–

can be proven to be unbiased and more efficient than $\widehat{\sigma}_N^2$ or $\tilde{\sigma}_N^2$.