Foundations of Machine Learning
CentraleSupélec — Fall 2017

6. Linear & logistic regressions

Chloé-Agathe Azencott
Centre for Computational Biology, Mines ParisTech
chloe-agathe.azencott@mines-paristech.fr

Learning objectives

- **Density estimation:**
  - Define **parametric methods.**
  - Define the **maximum likelihood estimator** and compute it for **Bernouilli**, **multinomial** and **Gaussian** densities.
  - Define the **Bayes estimator** and compute it for **normal priors.**
- **Supervised learning:**
  - Compute the maximum likelihood estimator / least-square fit solution for **linear regression.**
  - Compute the maximum likelihood estimator for **logistic regression.**

Density estimation

Parametric methods

- $\mathcal{X} = \{\boldsymbol{x}^i\}_{i=1,\dots,n}$ $\quad \boldsymbol{x}^i \sim p(\boldsymbol{x}|\theta)$
- **Parametric estimation:**
  - **assume a form for p(x|θ)**
    E.g. $p(x_j|\theta_j) \sim \mathcal{N}(\mu_j, \sigma_j^2)$ $\quad \theta = \{\mu_1, \sigma_1, \dots, \mu_p, \sigma_p\}$
  - Goal: estimate θ using $\mathcal{X}$
  - usually assume that $\boldsymbol{x}^i$ **independent and identically distributed** (iid)

Maximum likelihood estimation

- **Find θ such that $\mathcal{X}$ is the most likely to be drawn.**
- **Likelihood** of θ given the i.i.d. sample $\mathcal{X}$:
- **Log likelihood:**
- **Maximum likelihood estimation (MLE):**

Bernouilli density

- Two states: failure / success
  $x \in \{0, 1\}$
  $P(X = x|p_0) = {p_0}^x (1 - p_0)^{(1-x)}$
  $\mathcal{X} = \{x^i\}_{i=1,\ldots,n}$
- **MLE estimate of p_0**

Multinomial density

- Consider **K mutually exclusive and exhaustive classes**
  - Each class occurs with probability p_k $\sum_{k=1}^{K} p_k = 1$
  - x_1, x_2, …, x_K indicator variables: x_1=1 if the outcome is class k and 0 otherwise

$$P(x_1^i, x_2^i, \ldots, x_K^i) = \prod_{k=1}^{K} p_k^{x_k^i}$$

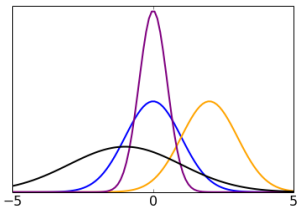
- The **MLE of p_k** is

$$\hat{p}_k = \frac{1}{n} \sum_{i=1}^{n} x_k^i$$

Gaussian distribution

- Gaussian distribution = normal distribution

$$x \sim \mathcal{N}(\mu, \sigma^2)$$



$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

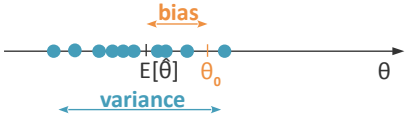
MLE estimates of μ and σ:

Bias-variance tradeoff

- **Mean squared error of the estimator:**

$$\begin{aligned} \mathrm{MSE}(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta_0)^2] \\ &= \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}^2(\hat{\theta}) \end{aligned}$$

A biased estimator may achieve better MSE than an unbiased one.



Bayes estimator

- Treat θ as a random variable with prior p(θ)
- **Bayes rule:** $p(\theta|\mathcal{X}) = \frac{p(\mathcal{X}|\theta)p(\theta)}{p(\mathcal{X})}$
- **Density estimation**

$$p(\boldsymbol{x}|\mathcal{X}) = \int p(\boldsymbol{x}, \theta|\mathcal{X})d\theta = \int p(\boldsymbol{x}|\theta)p(\theta|\mathcal{X})d\theta.$$

- **Maximum likelihood estimate (MLE):**

$$\theta_{\text{MLE}} = \arg\max_{\theta} p(\mathcal{X}|\theta)$$

- **Bayes estimate:**

$$\theta_{\text{Bayes}} = \arg\min_{\hat{\theta}} \mathbb{E}[L(\hat{\theta}, \theta)]$$

$$L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2 \quad \Rightarrow \quad \theta_{\text{Bayes}} = \mathbb{E}[\theta|\mathcal{X}] = \int \theta p(\theta|\mathcal{X})d\theta$$

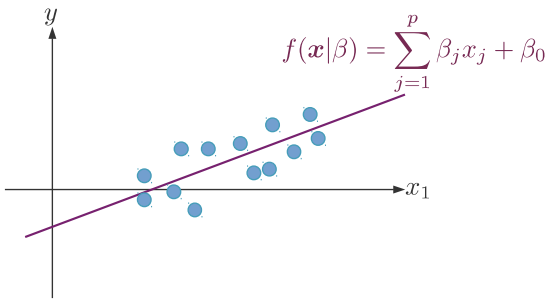
Bayes estimator: Normal prior

- n data points (iid) $x^i \sim \mathcal{N}(\theta, \sigma_0^2)$ $\quad\quad$ $\theta \sim \mathcal{N}(\mu, \sigma^2)$
- MLE of θ: $\hat{\theta}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x^i$

  Bayes estimator of θ: 

Linear regression

$\boldsymbol{x} \in \mathbb{R}^p, y \in \mathbb{R}$ $\quad\quad$ $\mathcal{D} = \{\boldsymbol{x}^i, y^i\}_{i=1,\ldots,n}$



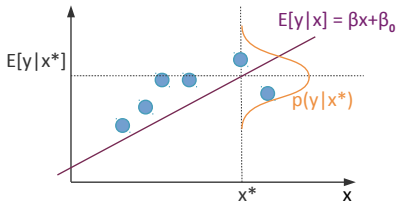
$$f(\boldsymbol{x}|\beta) = \sum_{j=1}^{p} \beta_j x_j + \beta_0$$

Linear regression: MLE

- Assume **error is Gaussian distributed**

$$y = g(\boldsymbol{x}) + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

- Replace g with its estimator f $f(\boldsymbol{x}|\beta) = \sum_{j=1}^{p} \beta_j x_j + \beta_0$



$$p(y|\boldsymbol{x}) \sim \mathcal{N}(f(\boldsymbol{x}|\beta), \sigma^2)$$

Gauss-Markov Theorem

- Under the assumption that $\epsilon \sim \mathcal{N}(0, \sigma^2)$

  the least-squares estimator of β is its (unique) best linear unbiased estimator.

- **Best Linear Unbiased Estimator (BLUE):**

  Var(β̂) < Var(β*) for any β* that is a linear unbiased estimator of β

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

$$\begin{aligned} Var(\hat{\beta}) &= \mathbb{E}[(X^\top X)^{-1} X^\top \epsilon\epsilon^\top X (X^\top X)^{-1}] \\ &= (X^\top X)^{-1} X^\top \sigma^2 I X (X^\top X)^{-1} \\ &= \sigma^2 (X^\top X)^{-1} \end{aligned}$$

$$\beta^* = Ay$$

$$Var(\beta^*) = \sigma^2 DD^\top + Var(\hat{\beta})$$

psd and minimal for D=0

$$D = A - (X^\top X)^{-1} X^\top$$

Correlated variables

- If the variables are **decorrelated**:
  - Each coefficient can be estimated separately;
  - **Interpretation** is easy:

    "A change of 1 in x_j is associated with a change of βj in Y, while everything else stays the same."
- **Correlations between variables cause problems:**
  - The **variance** of all coefficients tend to increase;
  - Interpretation is much harder

    when x_j changes, so does everything else.

Logistic regression

What about classification?

- **Model Pr(Y=1|X) as a linear function?** 

Maximum likelihood estimation of logistic regression coefficients

$$\log \frac{P(y=1|\boldsymbol{x})}{1-P(y=1|\boldsymbol{x})} = \beta^\top \boldsymbol{x} + \beta_0$$

- **Log likelihood for n observations:** $\mathcal{D} = \{\boldsymbol{x}^i, y^i\}_{i=1,\ldots,n}$

- **Gradient of the log likelihood:**

Summary

- **MAP estimate:**
$$\theta_{\text{MAP}} = \arg\max_\theta p(\theta|\mathcal{X})$$
- **MLE:**
$$\theta_{\text{MLE}} = \arg\max_\theta p(\mathcal{X}|\theta)$$
- **Bayes estimate:**
$$\theta_{\text{Bayes}} = \mathbb{E}[\theta|\mathcal{X}] = \int \theta p(\theta|\mathcal{X}) d\theta$$
- Assuming Gaussian error, maximizing the likelihood is equivalent to minimizing the RSS.
- **Linear regression MLE:**
$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$
- **Logistic regression MLE:** solve with gradient descent.

References

- *A Course in Machine Learning.* `http://ciml.info/dl/v0_99/ciml-v0_99-all.pdf`
  - **Least-squares regression:** Chap 7.6
- *The Elements of Statistical Learning.* `http://web.stanford.edu/~hastie/ElemStatLearn/`
  - **Least-squares regression:** Chap 2.2.1, 3.1, 3.2.1
  - **Gauss-Markov theorem:** Chap 3.2.3