Foundations of Machine Learning
CentraleSupélec — Fall 2016

5. Linear & logistic regressions

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

Practical matters

- *L'apprentissage artificiel. Concepts et algorithmes.* Antoine Cornuéjols et Laurent Miclet.

Learning objectives

- 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.**
- Compute the maximum likelihood estimator / least-square fit solution for **linear regression.**
- Compute the maximum likelihood estimator for **logistic regression.**

Parametric methods

Parametric methods

- $\mathcal{X} = \{\boldsymbol{x}^i\}_{i=1,\ldots,n}$ $\quad\quad$ $\boldsymbol{x}^i \sim p(\boldsymbol{x}|\theta)$
- **Parametric estimation:**
  - **assume a form for p(x|θ)**

    E.g. $p(x_j|\theta_j) \sim \mathcal{N}(\mu_j, \sigma_j^2)$ $\quad\quad$ $\theta = \{\mu_1, \sigma_1, \ldots, \mu_p, \sigma_p\}$
  - 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 $\mathcal{X}$:

$$\ell(\theta|\mathcal{X}) = p(\mathcal{X}|\theta) = p(\boldsymbol{x}^1|\theta)p(\boldsymbol{x}^2|\theta)\ldots p(\boldsymbol{x}^n|\theta)$$

- **Log likelihood:**

$$\mathcal{L}(\theta|\mathcal{X}) = \log \ell(\theta|\mathcal{X}) = \log p(\boldsymbol{x}^1|\theta) + \cdots + \log p(\boldsymbol{x}^n|\theta)$$

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

$$\theta^* = \arg\max_\theta \mathcal{L}(\theta|\mathcal{X})$$

Bernouilli density

- Two states: failure / success

  $x \in \{0, 1\}$

  $$\mathcal{X} = \{x^i\}_{i=1,\ldots,n}$$

  $P(X = x|p_0) = {p_0}^x (1 - p_0)^{(1-x)}$

  What is the MLE estimate of p_0?

Bernouilli density

- Two states: failure / success

  $x \in \{0, 1\}$

  $P(X = x|p_0) = {p_0}^x (1 - p_0)^{(1-x)}$

  $\mathcal{X} = \{x^i\}_{i=1,\ldots,n}$

  What is the MLE estimate of p_0?

- **Log likelihood:**

Bernouilli density

- Two states: failure / success

  $x \in \{0, 1\}$

  $$\mathcal{X} = \{x^i\}_{i=1,\ldots,n}$$

  $P(X = x|p_0) = {p_0}^x(1 - p_0)^{(1-x)}$

  What is the MLE estimate of p_0?

- **Log likelihood:**

$$L(p_0|\mathcal{X}) = \log P(\mathcal{X}|p_0) = \sum_{i=1}^{n} \left(x^i \log p_0 + (1 - x^i) \log(1 - p_0)\right)$$

- **To maximize the likelihood:** set its gradient to 0.

Bernouilli density

- Two states: failure / success

  $x \in \{0, 1\}$

  $\mathcal{X} = \{x^i\}_{i=1,\dots,n}$

  $P(X = x|p_0) = {p_0}^x (1 - p_0)^{(1-x)}$

  What is the MLE estimate of p_0?

- **Log likelihood:**

$$L(p_0|\mathcal{X}) = \log P(\mathcal{X}|p_0) = \sum_{i=1}^{n} \left(x^i \log p_0 + (1 - x^i) \log(1 - p_0) \right)$$

- **To maximize the likelihood:** set its gradient to 0.

$$\hat{p_0} = \frac{1}{n} \sum_{i=1}^{n} x^i$$

Multinomial density

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

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

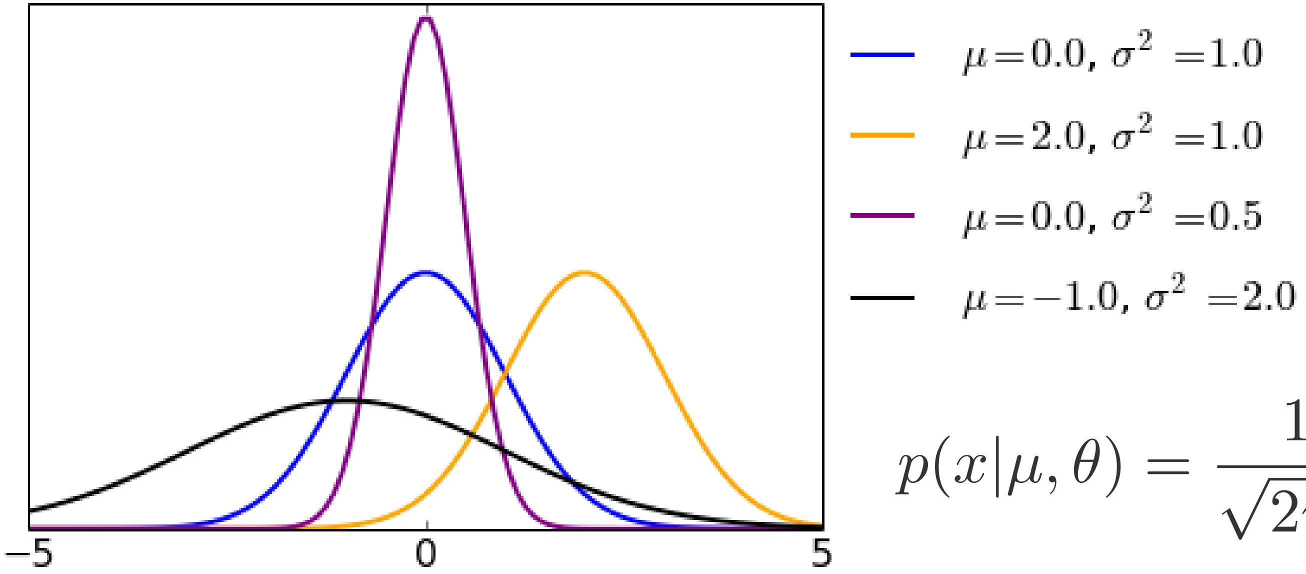
- The **MLE of p_k** is

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

Gaussian distribution

- Gaussian distribution = normal distribution

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



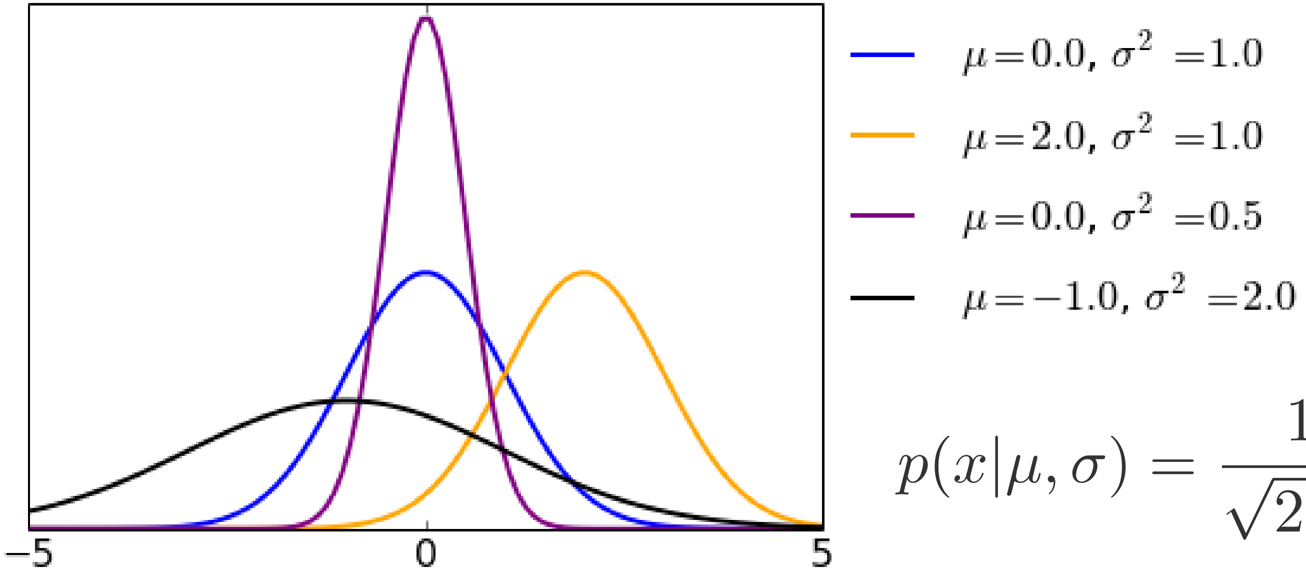
$$p(x|\mu, \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Compute the MLE estimates of μ and σ.

Gaussian distribution

- Gaussian distribution = normal distribution

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



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

Compute the MLE estimates of μ and σ.

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x^i$$

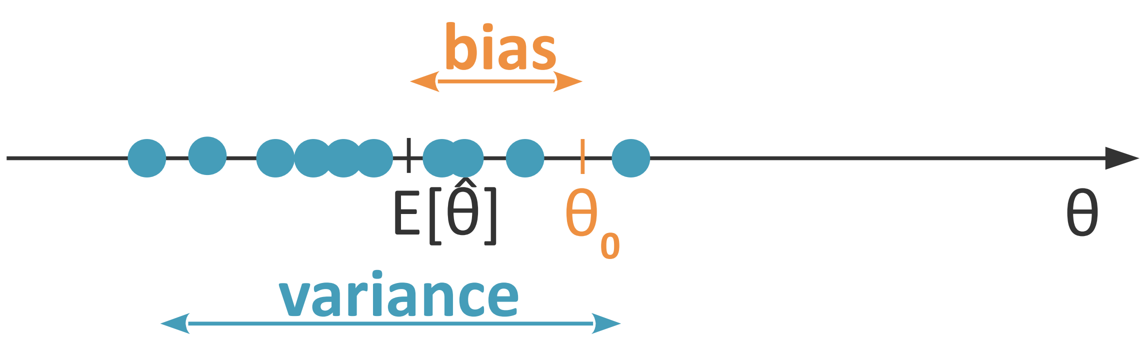
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x^i - \hat{\mu})^2$$

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

$$P(C|\boldsymbol{x}) = \frac{P(C)\,p(\boldsymbol{x}|C)}{p(\boldsymbol{x})}$$

prior

likelihood

posterior

evidence

- Treat θ as a random variable with prior p(θ)
- **Bayes rule:**

$$p(\theta|\mathcal{X}) = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\mathcal{X})}$$

- **Density estimation at x:**

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

Bayes estimator

- Treat θ as a random variable with prior p(θ)
- **Bayes rule:**

$$p(\theta|\mathcal{X}) = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\mathcal{X})}$$

- **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 a posteriori (MAP) estimate:**

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

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

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

- **Bayes estimate:**

$$\theta_{\text{Bayes}} = \mathbb{E}[\theta|\mathcal{X}] = \int \theta p(\theta|\mathcal{X})d\theta$$

Bayes estimator: Normal prior

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

  Compute the Bayes estimator of θ

Bayes estimator: Normal prior

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

Compute the Bayes estimator of θ $\theta_{\mathrm{Bayes}} = \mathbb{E}[\theta|\mathcal{X}]$

$$p(u|m,s) = \frac{1}{\sqrt{2\pi}s}\exp\left[-\frac{(u-m)^2}{2s^2}\right]$$

Hint:

Compute p(θ|$\mathcal{X}$) and show that it follows a normal distribution

Bayes estimator: Normal prior

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

Compute the Bayes estimator of θ $\theta_{\mathrm{Bayes}} = \mathbb{E}[\theta|\mathcal{X}]$

p(θ|$\mathcal{X}$) follows a normal distribution with

- mean

$$\frac{n\hat{\theta}_{\mathrm{MLE}}\sigma^2 + \mu\sigma_0^2}{n\sigma^2 + \sigma_0^2} = \frac{1/\sigma_0^2}{1/\sigma_0^2 + 1/n\sigma^2}\hat{\theta}_{\mathrm{MLE}} + \frac{1/\sigma^2}{n/\sigma_0^2 + 1/\sigma^2}\mu$$

- variance $\frac{\sigma^2\sigma_0^2}{n\sigma^2 + \sigma_0^2}$

$$p(\theta|\mathcal{X}) = \frac{1}{\sqrt{2\pi}s}\exp\left[-\frac{(\theta - m)^2}{2s^2}\right]$$

Bayes estimator: Normal prior

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

Compute the Bayes estimator of θ $\theta_{\text{Bayes}} = \boxed{\mathbb{E}[\theta|\mathcal{X}]}$

p(θ|$\mathcal{X}$) follows a normal distribution with

- mean

$$\frac{n\hat{\theta}_{\text{MLE}}\sigma^2 + \mu\sigma_0^2}{n\sigma^2 + \sigma_0^2} = \frac{1/\sigma_0^2}{1/\sigma_0^2 + 1/n\sigma^2}\hat{\theta}_{\text{MLE}} + \frac{1/\sigma^2}{n/\sigma_0^2 + 1/\sigma^2}\mu$$

- variance $\frac{\sigma^2\sigma_0^2}{n\sigma^2 + \sigma_0^2}$

$$p(\theta|\mathcal{X}) = \frac{1}{\sqrt{2\pi s}}\exp\left[-\frac{(\theta - \boxed{m})^2}{2s^2}\right]$$

Bayes estimator: Normal prior

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

$$\theta_{\text{Bayes}} = \frac{n/\sigma_0^2}{n/\sigma_0^2 + 1/\sigma^2} \underbrace{\boxed{\hat{\theta}_{\text{MLE}}}}_{\textbf{sample mean}} + \frac{1/\sigma^2}{n/\sigma_0^2 + 1/\sigma^2} \underbrace{\boxed{\mu}}_{\textbf{prior mean}}$$

Bayes estimator: Normal prior

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

$$\theta_{\text{Bayes}} = \boxed{\frac{n/\sigma_0^2}{n/\sigma_0^2 + 1/\sigma^2}}\,\boxed{\hat{\theta}_{\text{MLE}}} + \boxed{\frac{1/\sigma^2}{n/\sigma_0^2 + 1/\sigma^2}}\,\boxed{\mu}$$

sample mean

prior mean

large when σ is...

large when n is...

Bayes estimator: Normal prior

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

$$\theta_{\mathrm{Bayes}} = \frac{n/\sigma_0^2}{n/\sigma_0^2 + 1/\sigma^2}\hat{\theta}_{\mathrm{MLE}} + \frac{1/\sigma^2}{n/\sigma_0^2 + 1/\sigma^2}\mu$$

- When n ↗: θ_{Bayes} gets closer to the sample average (uses information from the sample).
- When σ is small, θ_{Bayes} gets closer to μ (little uncertainty about the prior).

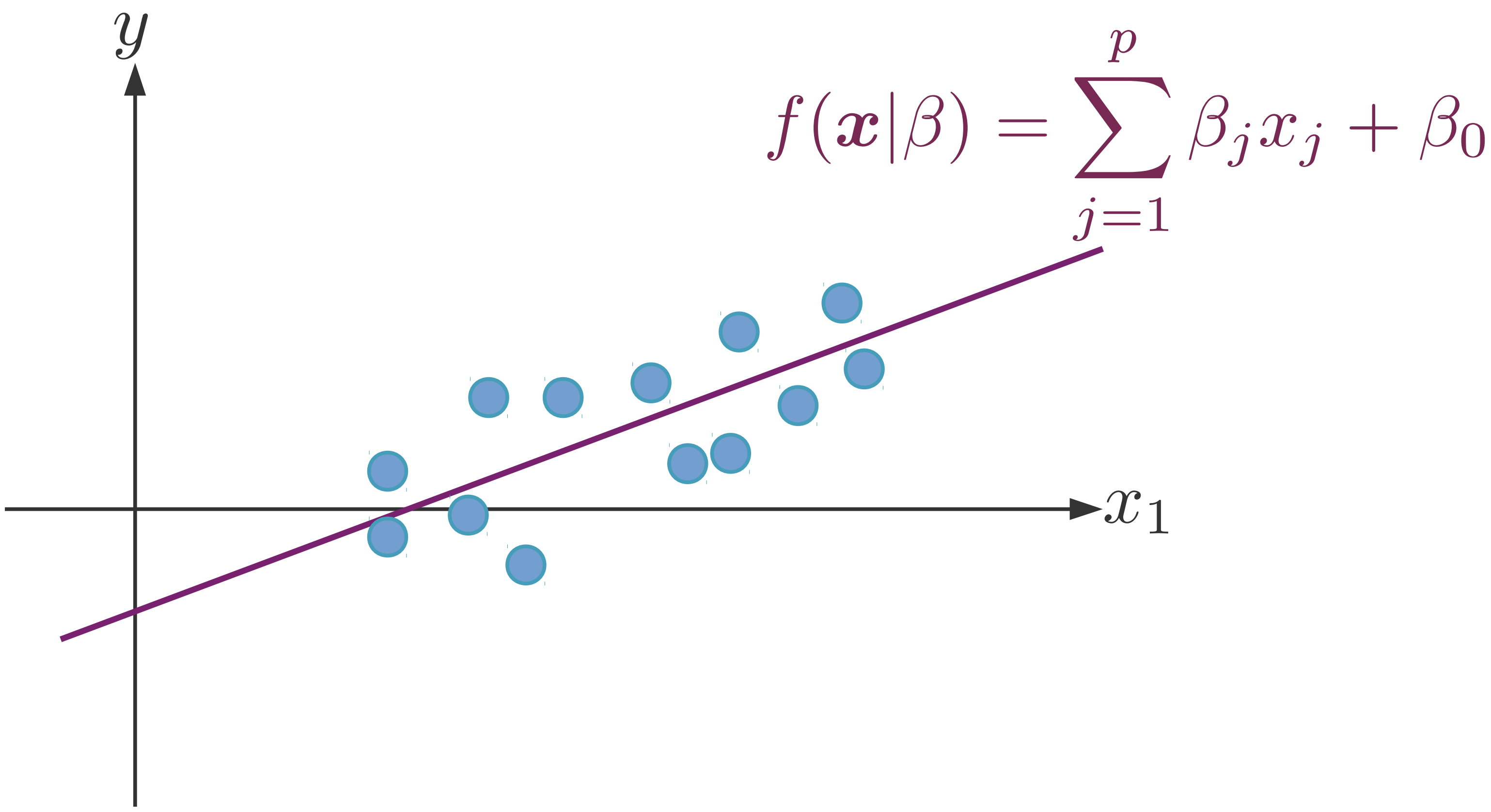
Linear regression

x_1

Linear regression

$\boldsymbol{x} \in \mathbb{R}^p, y \in \mathbb{R}$

$\mathcal{D} = \{\boldsymbol{x}^i, y^i\}_{i=1,\ldots,n}$

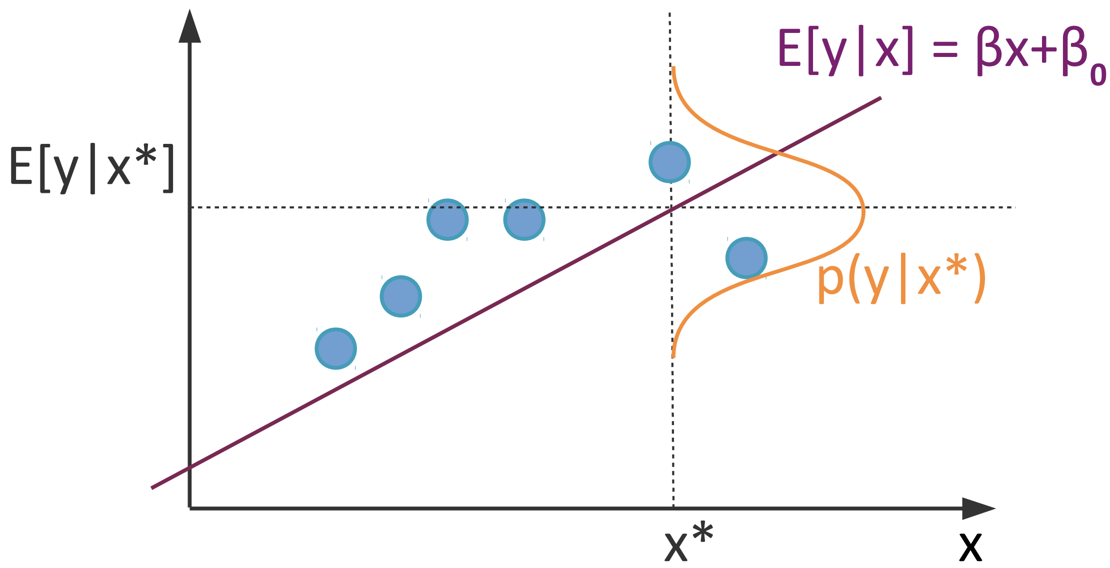


Linear regression: MLE

- Assume **error is Gaussian distributed**

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

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



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

MLE under Gaussian noise

- **Maximize (log) likelihood**

$$\mathcal{D} = \{\boldsymbol{x}^i, y^i\}_{i=1,\ldots,n}$$

$$\begin{aligned}\mathcal{L}(\beta|\mathcal{D}) &= \log \prod_{i=1}^n p(y^i|\boldsymbol{x}^i)p(\boldsymbol{x}^i) \\ &= \log \prod_{i=1}^n p(y^i|\boldsymbol{x}^i) + \boxed{\log \prod_{i=1}^n p(\boldsymbol{x}^i)}\end{aligned}$$

independent of β

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

$$\begin{aligned}\mathcal{L}(\beta|\mathcal{D}) &= \log\left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y^i - f(\boldsymbol{x}^i|\beta))^2}{2\sigma^2}\right] + \mathrm{Cte}\right) \\ &= \mathrm{Cte} - \frac{1}{2\sigma^2}\sum_{i=1}^n (y^i - f(\boldsymbol{x}^i|\beta))^2\end{aligned}$$

- **Assuming Gaussian error, maximizing the likelihood is equivalent to minimizing the sum of squared residuals.**

Linear regression least-squares fit

- Minimize the **residual sum of squares**

$$
\begin{aligned}
\mathrm{RSS}(\beta) &= \textstyle\sum_{i=1}^{n} (y^i - f(\boldsymbol{x}^i))^2 \\
&= \textstyle\sum_{i=1}^{n} \left(y^i - \beta_0 - \sum_{j=1}^{p} x_j^i \beta_j \right)^2 \\
&= (y - X\beta)^\top (y - X\beta)
\end{aligned}
$$

$$
X = \begin{pmatrix}
1 & x_1^1 & x_2^1 & \cdots & x_p^1 \\
1 & x_1^2 & x_2^2 & \cdots & x_p^2 \\
\vdots & \vdots & \ldots & \vdots & \\
1 & x_1^n & x_2^n & \cdots & x_p^n
\end{pmatrix}
$$

Linear regression least-squares fit

- Minimize the **residual sum of squares**

$$
\begin{aligned}
\mathrm{RSS}(\beta) &= \textstyle\sum_{i=1}^{n} (y^i - f(\boldsymbol{x}^i))^2 \\
&= \textstyle\sum_{i=1}^{n} \left(y^i - \beta_0 - \sum_{j=1}^{p} x_j^i \beta_j \right)^2 \\
&= (y - X\beta)^\top (y - X\beta)
\end{aligned}
$$

Historically:

- Carl Friedrich Gauss (to predict the location of Ceres)
- Adrien Marie Legendre

Linear regression least-squares fit

- Minimize the **residual sum of squares**

$$
\begin{aligned}
\mathrm{RSS}(\beta) &= \textstyle\sum_{i=1}^{n} (y^i - f(\boldsymbol{x}^i))^2 \\
&= \textstyle\sum_{i=1}^{n} \left(y^i - \beta_0 - \sum_{j=1}^{p} x_j^i \beta_j \right)^2 \\
&= (y - X\beta)^\top (y - X\beta)
\end{aligned}
$$

Estimate β. Under which condition is your estimate unique?

Linear regression least-squares fit

- Minimize the **residual sum of squares**

$$\begin{aligned}\mathrm{RSS}(\beta) &= \textstyle\sum_{i=1}^{n}(y^i - f(\boldsymbol{x}^i))^2 \\ &= \textstyle\sum_{i=1}^{n}\left(y^i - \beta_0 - \sum_{j=1}^{p} x_j^i \beta_j\right)^2 \\ &= (y - X\beta)^\top (y - X\beta)\end{aligned}$$

- Assuming **X has full column rank** (and hence $X^\top X$ invertible):

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

Linear regression least-squares fit

- Minimize the **residual sum of squares**

$$
\begin{aligned}
\mathrm{RSS}(\beta) &= \sum_{i=1}^{n}(y^i - f(\boldsymbol{x}^i))^2 \\
&= \sum_{i=1}^{n}\left(y^i - \beta_0 - \sum_{j=1}^{p} x_j^i \beta_j\right)^2 \\
&= (y - X\beta)^\top (y - X\beta)
\end{aligned}
$$

- Assuming **X has full column rank** (and hence $X^T X$ invertible):

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

- If X is rank-deficient, use a **pseudo-inverse.**

A **pseudo-inverse** of A
is a matrix G s. t. AGA = A

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.

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($\hat{\beta}$) < Var(β*) for any β* that is a linear unbiased estimator of β

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($\hat{\beta}$) < Var(β*) for any β* that is a linear unbiased estimator of β

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

$$\beta^* = Ay$$

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($\hat{\beta}$) < Var(β*) for any β* that is a linear unbiased estimator of β

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

$$\mathbb{E}[\beta^*] = \beta$$

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

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($\hat{\beta}$) < Var(β*) for any β* that is a linear unbiased estimator of β

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

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

$$\beta^* = Ay$$

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

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

psd and minimal
for D=0

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?

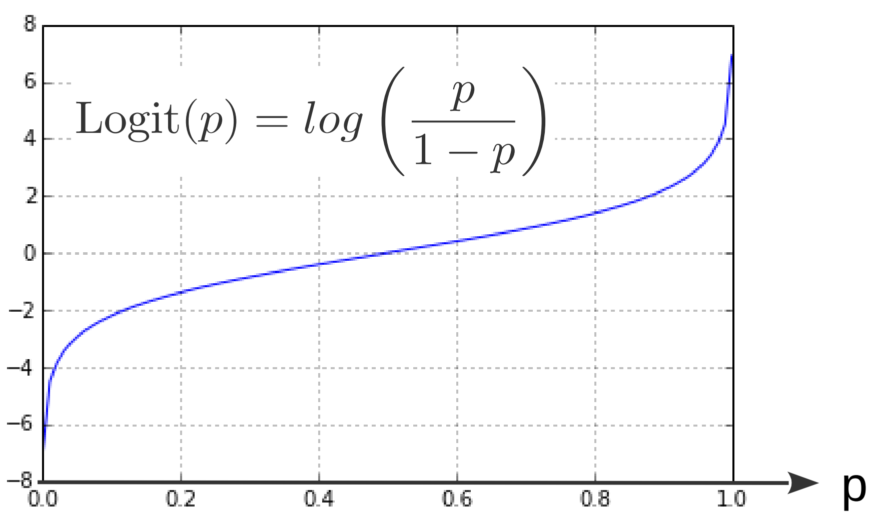
What about classification?

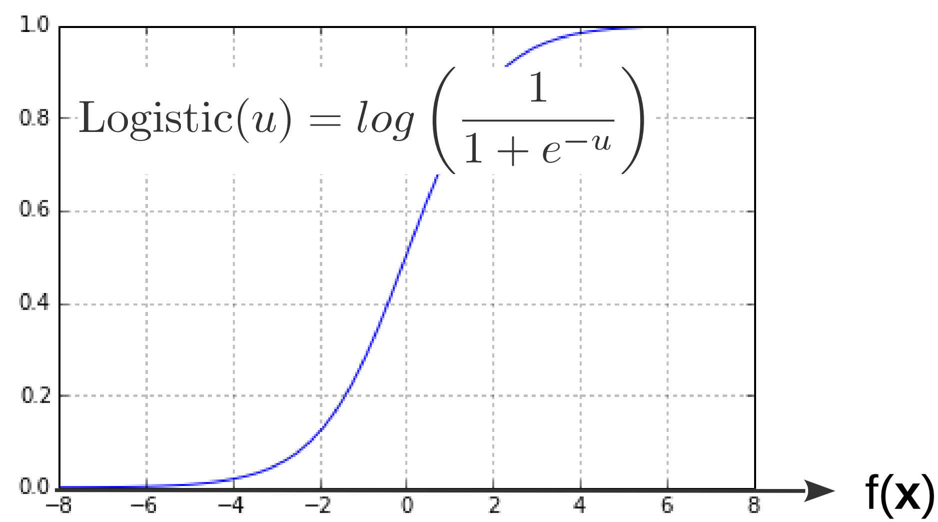
- Model Pr(Y=1|X) as a linear function?

  Problems?

What about classification?

- **Model P(Y=1|x) as a linear function?**
  - Problem: P(Y=1|**x**) **must be between 0 and 1.**
  - **Non-linearity:**
    - If P(Y=1|**x**) close to +1 or 0, **x** must change a lot for y to change;
    - If P(Y=1|**x**) close to 0.5, that's not the case.
  - Hence: use a **logit transformation**





→ **Logistic regression.**

Maximum likelihood estimation of logistic regression coefficients

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

- **Compute the log likelihood for n observations**

$$\mathcal{D} = \{\boldsymbol{x}^i, y^i\}_{i=1,\ldots,n}$$

Maximum likelihood estimation of logistic regression coefficients

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

- **Compute the log likelihood for n observations**

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

$$\begin{array}{ccc} \boldsymbol{x} & \leftarrow & [1, x_1, \ldots, x_p] \\ \beta & \leftarrow & [\beta_0, \beta_1, \ldots, \beta_p] \end{array}$$

$$\begin{array}{rcl} \mathcal{L}(\beta|\mathcal{D}) & = & \sum_{i=1}^n \log P(y^i|\boldsymbol{x}^i) + \mathrm{Cte} \\ & = & \sum_{i=1}^n \big(y^i \log g^i + (1-y^i)\log(1-g^i)\big) \end{array}$$

Maximum likelihood estimation of logistic regression coefficients

$$\mathcal{L}(\beta|\mathcal{D}) = \sum_{i=1}^{n} \left(y^i \log g^i + (1 - y^i) \log(1 - g^i) \right)$$

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

$$\begin{aligned} \boldsymbol{x} &\leftarrow [1, x_1, \ldots, x_p] \\ \beta &\leftarrow [\beta_0, \beta_1, \ldots, \beta_p] \end{aligned}$$

- **Compute the gradient of the log likelihood** $\nabla_\beta \mathcal{L}$

Maximum likelihood estimation of logistic regression coefficients

$$\mathcal{L}(\beta|\mathcal{D}) = \sum_{i=1}^{n} \left(y^i \log g^i + (1-y^i)\log(1-g^i)\right)$$

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

$$\begin{aligned} \boldsymbol{x} &\leftarrow [1, x_1, \ldots, x_p] \\ \beta &\leftarrow [\beta_0, \beta_1, \ldots, \beta_p] \end{aligned}$$

- **Compute the gradient of the log likelihood** $\nabla_\beta \mathcal{L}$

$$\nabla_\beta g^i = \boldsymbol{x}^i g^i (1-g^i)$$

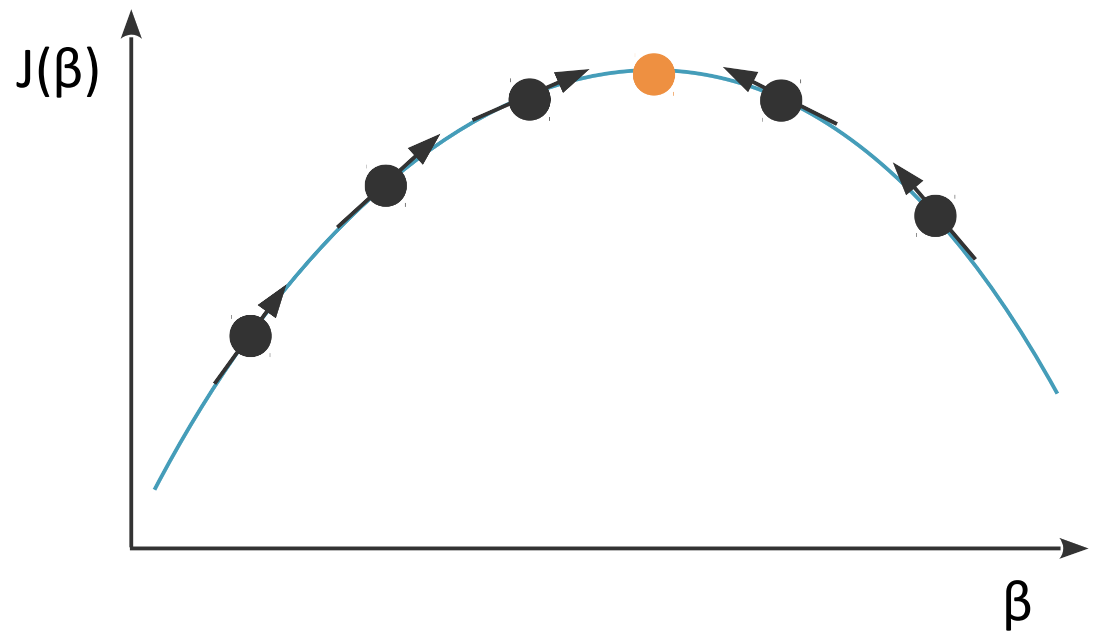
$$\nabla_\beta \mathcal{L} = \sum_{i=1}^{n} (y^i - g^i)\boldsymbol{x}^i$$

- **To maximize the likelihood:**
  - set the gradient to 0 $$\sum_{i=1}^{n}\left(y^i - \frac{1}{1+e^{-\beta^\top \boldsymbol{x}^i}}\right)\boldsymbol{x}^i = 0$$
  - cannot be solved analytically
  - L **concave** so we can use **gradient ascent** (no local minima)

Gradient ascent

- J **concave** in β
- **Update rule:**

$$\beta^{(t+1)} \leftarrow \beta^{(t)} + \eta \nabla_\beta J(\beta^{(t)})$$



- η > 0: **learning rate**
- Iterate until change < ε
- Other methods
  - Newton methods, conjugate gradient ascent, IRLS.

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 ascent.

kaggle challenge project

How Many Bikes? Challenge

`https://www.kaggle.com/c/how-many-bikes`

- **Predict the number of shared bikes** that are rented in an American city
  - Regression
  - From weather, holiday, date & time.
- **Evaluation on**
  - Insights learned
  - Prediction performance.

Evaluation

- **Kaggle project (30 pts)** **December 16, 2016**
  - Written report (25 pts)
    - Evaluate methods from 5 families (cross-validation / leaderboard)
    - Features pre-processing
    - Choice of final models
    - Additional insights, models, ideas
  - Position in the leaderboard (5pts)