Introduction to Machine Learning
NPFL 054

http://ufal.mff.cuni.cz/course/npfl054

Barbora Hladká **Martin Holub**

{Hladka | Holub}@ufal.mff.cuni.cz

Charles University,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

Maximum Likelihood Estimation (MLE)

Example

The binomial distribution is the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success with probability p, $X \sim Bin(n, p)$.

Probabilistic mass function $\Pr(X = k) = f(k; n, p) = \frac{n!}{k!(n-k)!} p^k (1-p)^{(n-k)}$

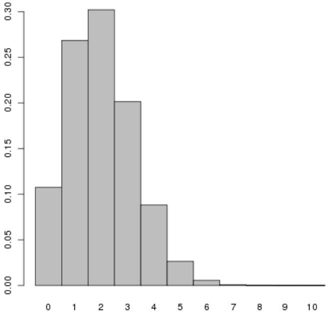
Coin tossing

Let $n = 10$, x represents the number of successes in 10 trials and probability of head on one trial is p. Then

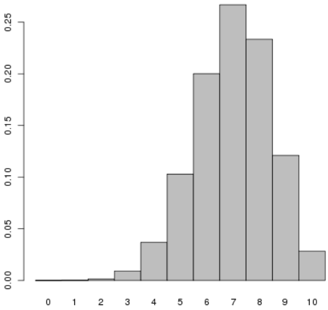
$$f(x; 10, p) = \frac{10!}{x!(10-x)!} p^x (1-p)^{(10-x)}$$

Example

$f(x; 10, p = 0.2)$



$f(x; 10, p = 0.0.7)$



MLE

- $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$

Assumption

$\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent and identically distributed with an unknown probability density function $f(\mathbf{X}; \mathbf{\Theta})$

- unknown parameters $\mathbf{\Theta}$
- joint density function $f(\mathbf{x}_1, \ldots, \mathbf{x}_n; \mathbf{\Theta}) \overset{i.i.d.}{=} \prod_{i=1}^{n} f(\mathbf{x}_i; \mathbf{\Theta})$

We determine what value of $\mathbf{\Theta}$ would make the data that we observed most likely.

MLE

MLE is a method for estimating population parameters from data.

Goal: identify the population that is most likely to have generated the sample.

Likelihood function

$$\mathcal{L}(\Theta|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} f(\mathbf{x}_i; \boldsymbol{\Theta}) \quad (1)$$

Log-likelihood function

$$\log \mathcal{L}(\Theta|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i=1}^{n} \log f(\mathbf{x}_i; \boldsymbol{\Theta}) \quad (2)$$

Maximum likelihood estimate of Θ

$$\Theta^{\star}_{MLE} = \mathrm{argmax}_{\boldsymbol{\Theta}} \log \mathcal{L}(\boldsymbol{\Theta}|\mathbf{x}_1, \ldots, \mathbf{x}_n) \quad (3)$$

MLE

Analytically

- Likelihood equation: $\frac{\partial \log \mathcal{L}(\Theta|\mathbf{X})}{\partial \Theta_i} = 0$ at Θ_i for all $i = 1, \ldots, m$
- Maximum, not minimum: $\frac{\partial^2 \mathcal{L}(\theta|\mathbf{X})}{\partial \Theta_i^2} < 0$

In practice, it is usually not possible to obtain an analytic solution (many parameters, probability density function is highly non-liner).

Numerically

- Use an optimization algorithm (for ex. Gradient Descent)

MLE
Binomial distribution

Estimate the probability p that a coin lands head using the result of n coin tosses, k of which resulted in heads.

- $f(k; n, p) = \frac{n!}{k!(n-k)!}(p)^k(1-p)^{(n-k)}$
- $\mathcal{L}(p|n, k) = \frac{n!}{k!(n-k)!}(p)^k(1-p)^{(n-k)}$
- $\log \mathcal{L}(p|n, k) = \log \frac{n!}{k!(n-k)!} + k \log p + (n-k) log(1-p)$
- $\frac{\partial \log \mathcal{L}(p|n,k)}{\partial p} = \frac{k}{p} + \frac{n-k}{1-p} = 0$
- $p^{\star}_{MLE} = \frac{k}{n}$

MLE
Logistic regression

Logistic regression models conditional probability using linear function.

$$h(\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{\Theta}^T \mathbf{x}}} = \Pr(y = 1|\mathbf{x})$$

Learn $\mathbf{\Theta}^\star$ from $Data = \{\langle \mathbf{x}_i, y_i \rangle, y_i \in \{0, 1\}, i = 1, ..., n\}$.

Use MLE.

MLE
Logistic regression

$$h(\mathbf{x}; \mathbf{\Theta}) = \Pr(y = 1|\mathbf{x})$$

$$\prod_{i=1}^{n} \Pr(y = y_i|\mathbf{x}_i) = \prod_{i=1}^{n} h(\mathbf{x}_i; \mathbf{\Theta})^{y_i}(1 - h(\mathbf{x}_i; \mathbf{\Theta}))^{1-y_i}$$

$$\mathcal{L}(\mathbf{\Theta}|Data) = \prod_{i=1}^{n} h(\mathbf{x}_i; \mathbf{\Theta})^{y_i}(1 - h(\mathbf{x}_i; \mathbf{\Theta}))^{1-y_i}$$

$$\log \mathcal{L}(\mathbf{\Theta}|Data) = \sum_{i=1}^{n} y_i \log h(\mathbf{x}_i; \mathbf{\Theta}) + (1 - y_i) \log(1 - h(\mathbf{x}_i; \mathbf{\Theta}))$$

$$\mathbf{\Theta}^{\star}_{MLE} = \mathrm{argmax}_{\mathbf{\Theta}} \sum_{i=1}^{n} y_i \log h(\mathbf{x}_i; \mathbf{\Theta}) + (1 - y_i) \log(1 - h(\mathbf{x}_i; \mathbf{\Theta}))$$

MLE
Naïve Bayes classifier

$$\hat{y} = argmax_{y_k \in Y} \Pr(y_k) \prod_{j=1}^{m} \Pr(x_j | y_k)$$

MLE
Naïve Bayes classifier

Categorical features

- $\Theta_j(x|y) = \Pr(x|y)$, $x \in A_j, y \in Y$
- $\Theta(y) = \Pr(y)$, $y \in Y$

Where to get $\Theta_j(x|y)$ and $\Theta(y)$?

Learn them from $Data = \{\langle \mathbf{x}_i, y_i \rangle, y_i \in \mathcal{R}, i = 1, ..., n\}$.

Use MLE.

Theorem

The Maximum likelihood estimates for NB take the form

- $\Theta(y) = \frac{c_y}{n}$ *where* $c_y = \sum_{i=1}^{n} \delta(y_i, y)$
- $\Theta_j(x|y) = \frac{c_{j_{x|y}}}{c_y}$ *where* $c_{j_{x|y}} = \sum_{i=1}^{n} \delta(y_i, y)\delta(\mathbf{x}_{i_j}, x)$

MLE
Naïve Bayes classifier

Continuous features

Typical assumption: each continuous feature has a Gaussian distribution.

Theorem

The ML estimates for NB take the form

- $\overline{\mu_k} = \frac{\sum_{j=1}^{n} x_i^j \delta(Y^j = y_k)}{\sum_{j=1}^{n} \delta(Y^j = y_k)}$
- $\overline{\sigma_k}^2 = \frac{\sum_j (x_i^j - \overline{\mu_k})^2 \delta(Y^j = y_k)}{\sum_j \delta(Y^j = y_k)}$

- $\Theta_j(x|y_k) = \frac{1}{\sqrt{2\pi\overline{\sigma_k}^2}} e^{\frac{-(x-\overline{\mu_k})^2}{2\overline{\sigma_k}^2}}$

Least squares

Least squares

- seeking the parameter values that provide *most accurate* description of the data
- $\mathbf{\Theta}^\star = \mathrm{argmin}_{\mathbf{\Theta}} \sum_{i=1}^n (h(\mathbf{x}_i) - y_i)^2$

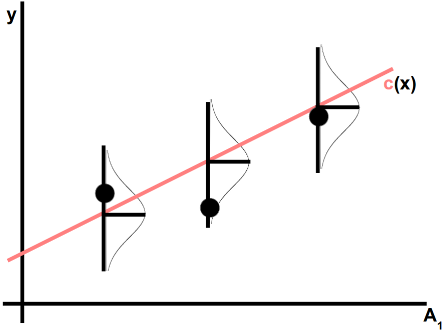
MLE

- seeking the parameter values that are *most likely* to have produced the data

Least squares

At each value of A_1, the output value y is subject to random error ϵ that is normally distributed $N(0, \sigma^2)$.

$y_i = \mathbf{\Theta}^T \mathbf{x}_i + \epsilon_i$



Least squares

- probability density function of the Normal distribution

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

- $$\mathcal{L}(\mu,\sigma|\epsilon) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\sum_{i=1}^{n}(\epsilon_i-\mu)^2}{2\sigma^2}}$$

- $\epsilon_i = y_i - \mathbf{\Theta}^T \mathbf{x}_i \sim N(0,\sigma^2)$. The likelihood distribution function is

$$\mathcal{L}(\mathbf{\Theta},\sigma|Data) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mathbf{\Theta}^T\mathbf{x}_i)^2}{2\sigma^2}}$$

Least squares

$$\log \mathcal{L}(\mathbf{\Theta}, \sigma | Data) = \sum_{i=1}^{n} [\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(y_i - \mathbf{\Theta}^T \mathbf{x}_i)^2}{2\sigma^2}]$$

$$\mathrm{argmax}_{\mathbf{\Theta}} \log \mathcal{L}(\mathbf{\Theta}, \sigma | Data) = \mathrm{argmax}_{\mathbf{\Theta}} \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(y_i - \mathbf{\Theta}^T \mathbf{x}_i)^2$$

$$\mathrm{argmax}_{\mathbf{\Theta}} \log \mathcal{L}(\Theta, \sigma | Data) = \mathrm{argmin}_{\mathbf{\Theta}} \sum_{i=1}^{n} (y_i - \mathbf{\Theta}^T \mathbf{x}_i)^2$$

The minimum least square estimates are equivalent to the maximum likelihood estimates under the assumption that Y is generated by adding random noise to the true target values characterized by the Normal distribution $N(0, \sigma^2)$.