Maximum Likelihood

Chris Williams

School of Informatics, University of Edinburgh

October 2008

Overview

- Maximum likelihood parameter estimation
- Example: multinomial
- Example: Gaussian
- ML parameter estimation in belief networks
- Properties of ML estimators
- Reading: Bishop §2.2 (multinomial), §2.3.4 (Gaussian)

Setting parameters

- We choose a parametric model $p(\mathbf{x}|\boldsymbol{\theta})$
- We are given data $\mathbf{x}_1, \ldots, \mathbf{x}_n$
- How can we choose $\boldsymbol{\theta}$ to best approximate the true density $p(\mathbf{x})$?
- Define the *likelihood* of $\mathbf{x}_i$ as

$$\mathcal{L}_i(\boldsymbol{\theta}) = p(\mathbf{x}_i|\boldsymbol{\theta})$$

- For points generated *independently and identically distributed* (iid) from $p(\mathbf{x})$, the likelihood of the data set is

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta})$$

- Often convenient to take logs,

$$L = \log \mathcal{L} = \sum_{i=1}^{n} \log p(\mathbf{x}_i|\boldsymbol{\theta})$$

- *Maximum likelihood* parameter estimation chooses $\boldsymbol{\theta}$ to maximize L (same as maximizing $\mathcal{L}$ as log is monotonic)

Example: multinomial distribution

- Consider an experiment with n independent trials
- Each trial can result in any of r possible outcomes (e.g. a die)
- θ_i denotes the probability of outcome i, $\sum_{i=1}^{r} \theta_i = 1$
- n_i denotes the number of trials resulting in outcome i, $\sum_{i=1}^{r} n_i = n$
- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$, $\mathbf{n} = (n_1, \ldots, n_r)$
- Show that

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{r} \theta_i^{n_i}$$

- Hence show that the maximum likelihood estimate for θ_i is

$$\hat{\theta}_i = \frac{n_i}{n}$$

Gaussian example

- likelihood for one data point x_i in 1-d

$$p(x_i|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp - \left\{ \frac{(x_i - \mu)^2}{2\sigma^2} \right\}$$

- Log likelihood for n data points

$$L = -\frac{1}{2} \sum_{i=1}^{n} \left[\log(2\pi\sigma^2) + \frac{(x_i - \mu)^2}{\sigma^2} \right]$$

- Show that

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

- and

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

- For the multivariate Gaussian

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

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$

ML parameter estimation in fully observable belief networks

$$P(X_1, \ldots, X_k|\boldsymbol{\theta}) = \prod_{j=1}^{k} P(X_j|Pa_j, \theta_j)$$

- Show that parameter estimation for θ_j depends only on statistics of (X_j, Pa_j)
- Discrete variables: CPTs

$$\hat{\theta}(X_2 = s_k|X_1 = s_j) = \frac{n_{jk}}{\sum_l n_{jl}}$$

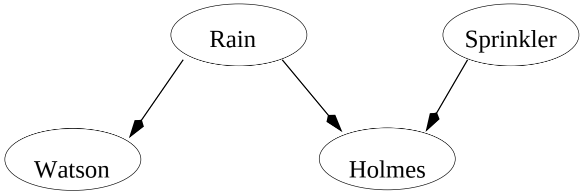
- Gaussian variables

$$Y = \mu_y + w_y(X - \mu_x) + N_y$$

Estimation of μ_x, μ_y, w_y and v_{N_y} is a linear regression problem (see Bishop §3.1.1)

Example of ML Learning in a Belief Network

R	S	H	W
n	n	n	n
n	n	n	n
y	n	y	y
n	n	n	n
n	n	n	n
n	n	n	y
n	n	n	n
n	n	n	y
n	n	n	n
y	y	y	y



From the table of data we obtain the following ML estimates for the CPTs

$$
\begin{aligned}
\hat{\theta}(R = yes) &= 2/10 = 0.2 \\
\hat{\theta}(S = yes) &= 1/10 = 0.1 \\
\hat{\theta}(W = yes|R = yes) &= 2/2 = 1 \\
\hat{\theta}(W = yes|R = no) &= 2/8 = 0.25 \\
\hat{\theta}(H = yes|R = yes, S = yes) &= 1/1 = 1.0 \\
\hat{\theta}(H = yes|R = yes, S = no) &= 1/1 = 1.0 \\
\hat{\theta}(H = yes|R = no, S = yes) &= 0/0 \\
\hat{\theta}(H = yes|R = no, S = no) &= 0/8 = 0.0
\end{aligned}
$$

Properties of ML estimators

- An estimator is **consistent** if it converges to the true value as the sample size $n \to \infty$. Consistency is a "good thing"
- **Bias**
  An estimator $\hat{\theta}$ is unbiased if $E[\hat{\theta}] = \theta$. The expectation is wrt data drawn from the model $p(\cdot|\theta)$
- The estimator $\hat{\mu}$ for the mean of a Gaussian is unbiased
- The estimator $\hat{\sigma}^2$ for the variance of a Gaussian is biased, with $E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2$
- For n very large ML estimators are approximately unbiased

- **Variance**
  One can also be interested in the variance of an estimator, i.e. $E[(\hat{\theta} - \theta)^2]$
- ML estimators have variance nearly as small as can be achieved by any estimator
- The MLE is approximately the minimum variance unbiased estimator (MVUE) of θ