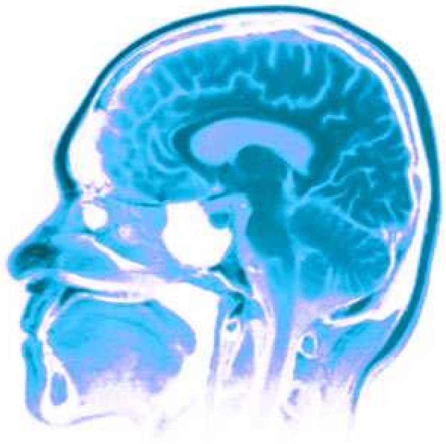
CPSC540



Maximum Likelihood

Nando de Freitas
2011
KPM Book Sections: *3.2.1, 3.2.2, 3.4, 3.5, 3.6*

MLE - definition

- The idea of **Maximum Likelihood Estimation (MLE)** is to find the parameters $\boldsymbol{\theta}$ that maximize the probability of the data $\mathcal{D}$ given these parameters:
$$\hat{\boldsymbol{\theta}} := \arg\max_{\boldsymbol{\theta}} \log p(\mathcal{D}|\boldsymbol{\theta})$$
- MLE assumes that the data has been generated by a distribution $p(\mathcal{D}|\boldsymbol{\theta}_0)$ for some true parameter $\boldsymbol{\theta}_0$.

MLE - properties

- For independent and identically distributed (i.i.d.) data from $p(x|\boldsymbol{\theta}_0)$, the MLE minimizes the **Kullback-Leibler divergence**:

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p(x_i|\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \tfrac{1}{N} \sum_{i=1}^{N} \log p(x_i|\boldsymbol{\theta}) - \tfrac{1}{N} \sum_{i=1}^{N} \log p(x_i|\boldsymbol{\theta}_0) \\
&= \arg\max_{\boldsymbol{\theta}} \tfrac{1}{N} \sum_{i=1}^{N} \log \frac{p(x_i|\boldsymbol{\theta})}{p(x_i|\boldsymbol{\theta}_0)} \\
&\longrightarrow \arg\min_{\boldsymbol{\theta}} \int \log \frac{p(x_i|\boldsymbol{\theta}_0)}{p(x_i|\boldsymbol{\theta})} p(x|\boldsymbol{\theta}_0) dx
\end{aligned}$$

MLE - properties

- Under smoothness and identifiability assumptions, the MLE is **consistent**:

$$\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$$

or equivalently,

$$\text{plim}(\hat{\boldsymbol{\theta}}) = \theta_0$$

or equivalently,

$$\lim_{N \to \infty} P(|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| > \alpha) \to 0$$

for every α.

MLE - properties

- The MLE is **asymptotically normal**. That is, as $N \to \infty$, we have:

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \Longrightarrow N(0, I^{-1})$$

  where I is the **Fisher Information matrix**.

- It is asymptotically optimal or **efficient**. That is, asymptotically, it has the lowest variance among all well behaved estimators. In particular it attains a lower bound on the CLT variance known as the **Cramer-Rao lower bound**.

- But what about issues like robustness and computation? Is MLE always the right option?

Bias and variance

- Note that the estimator is a function of the data: $\hat{\boldsymbol{\theta}} = g(\mathcal{D})$.
- Its **bias** is:

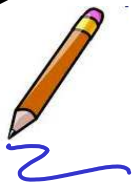
$$bias(\hat{\boldsymbol{\theta}}) = \mathbb{E}_{p(\mathcal{D}|\boldsymbol{\theta}_0)}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}_0 = \bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$$

- Its **variance** is:

$$\mathbb{V}(\hat{\boldsymbol{\theta}}) = \mathbb{E}_{p(\mathcal{D}|\boldsymbol{\theta}_0)}(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})^2$$

- Its **mean squared error** is:

$$\text{MSE} = \mathbb{E}_{p(\mathcal{D}|\boldsymbol{\theta}_0)}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = (\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^2 + \mathbb{E}_{p(\mathcal{D}|\boldsymbol{\theta}_0)}(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})^2$$



MLE for the univariate Gaussian distribution

- In the case of iid data sampled from a univariate Gaussian, the log-likelihood is given by

$$\ell(\mu, \sigma^2) \quad = \quad \sum_{i=1}^{N} \log \mathcal{N}(x_i|\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2 - \frac{N}{2}\ln \sigma^2 - \frac{N}{2}\ln(2\pi)$$

- To find the maximum of this function, we set the partial derivatives to 0 and solve. (We should check that the second derivative is positive.)

$$\begin{aligned} \hat{\mu} \quad &= \quad \frac{1}{N}\sum_{i=1}^{N} x_i = \overline{x} \\ \hat{\sigma}^2 \quad &= \quad \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - (\overline{x})^2 \end{aligned}$$

- The quantities $\sum_i x_i$, $\sum_i x_i^2$ and N are called the **sufficient statistics** of the data, since they capture all the relevant information needed for estimating the parameter.

MLE for the Bernoulli distribution

- We toss a coin N times and record the sequence of heads and tails, $\mathcal{D} = (x_1, x_2, \ldots, x_N)$. How do we estimate the probability of heads from this?

- The log-likelihood is given by

$$\ell(\theta) = \sum_{i=1}^{N} \log \mathrm{Ber}(x_i|\theta) = \sum_{i=1}^{N} \log \left[\theta^{x_i}(1-\theta)^{1-x_i}\right] = N_1 \log \theta + N_2 \log(1-\theta)$$

  where $N_1 = \sum_i x_i$ is the number of heads and $N_2 = \sum_i (1 - x_i)$ is the number of tails (these are the sufficient statistics).

- To find the MLE, we find the maximum of this expression as follows:

$$\begin{aligned} \frac{d\ell}{d\theta} &= \frac{N_1}{\theta} - \frac{N_2}{1-\theta} = 0 \\ \hat{\theta} &= \frac{N_1}{N_1 + N_2} \end{aligned}$$

  where $N_1 + N_2 = N$.

MLE for binary logistic regression

- Recall that binary logistic regression has the form

$$p(y_i|\mathbf{x}_i, \mathbf{w}) = \text{Ber}(y_i|\text{sigm}(\mathbf{w}^T\mathbf{x}_i))$$

where $\text{sigm}(\eta) = 1/(1 + e^{-\eta})$ and $y_i \in \{0, 1\}$. Let $\pi_i = \text{sigm}(\mathbf{w}^T\mathbf{x}_i)$.

- Then the negative log-likelihood of all the data is given by

$$\begin{aligned} J(\mathbf{w}) &= -\sum_{i=1}^{N} \log[\pi_i^{\mathbb{I}(y_i=1)} \times (1 - \pi_i)^{\mathbb{I}(y_i=0)}] \\ &= -\sum_{i=1}^{N} [y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)] \end{aligned}$$

This is also called the **cross-entropy** error function.

MLE for binary logistic regression

- The gradient and Hessian of $J(\mathbf{w})$ are given by:

$$\begin{aligned}
\mathbf{g}(\mathbf{w}) &= \frac{d}{d\mathbf{w}} J(\mathbf{w}) = \sum_i (\pi_i - y_i)\mathbf{x}_i = \mathbf{X}^T(\boldsymbol{\pi} - \mathbf{y}) \\
\mathbf{H} &= \frac{d}{d\mathbf{w}} \mathbf{g}(\mathbf{w})^T = \sum_i (\nabla_{\mathbf{w}} \pi_i)\mathbf{x}_i^T = \sum_i \pi_i(1-\pi_i)\mathbf{x}_i\mathbf{x}_i^T = \mathbf{X}^T \mathrm{diag}(\pi_i(1-\pi_i))\mathbf{X}
\end{aligned}$$

- One can show that $\mathbf{H}$ is positive definite; hence the NLL is **convex** and has a unique global minimum.

- To find this minimum, we will however have to learn a few things about optimization.

Revision

- Let $\mathbf{x}$ be an n-dimensional vector, and $f(\mathbf{x})$ a scalar-valued function. The gradient vector of f with respect to $\mathbf{x}$ is the following vector:

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

Derivative of a scalar product

$$\nabla_{\mathbf{x}} \; \mathbf{a}^T \mathbf{x} = \mathbf{a}$$

Derivative of a quadratic form

$$\nabla_{\mathbf{x}} \; \mathbf{x}^T \mathbf{A} \mathbf{x} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$$

If $\mathbf{A}$ is symmetric, this becomes $\nabla_{\mathbf{x}} \mathbf{x}^T \mathbf{A} \mathbf{x} = 2\mathbf{A}\mathbf{x}$.

- The **Hessian** matrix of a scalar valued function with respect to $\mathbf{x}$, written $\nabla^2_{\mathbf{x}} f(\mathbf{x})$ or simply as $\mathbf{H}$, is the $n \times n$ matrix of partial derivatives,

$$\nabla^2_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{bmatrix}$$

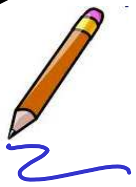
  Obviously this is a symmetric matrix, since

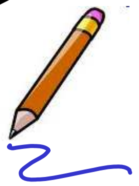
$$\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} = \frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_i}.$$

- We can think of the Hessian as the gradient of the gradient, which can be written as

$$\mathbf{H} = \nabla_{\mathbf{x}}(\nabla^T_{\mathbf{x}} f(\mathbf{x}))$$

- For the quadratic form $f(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x}$, the gradient vector is $(\mathbf{A} + \mathbf{A}^T)\mathbf{x}$, so the Hessian is $\mathbf{A} + \mathbf{A}^T$. If $\mathbf{A}$ is symmetric, this is $2\mathbf{A}$.

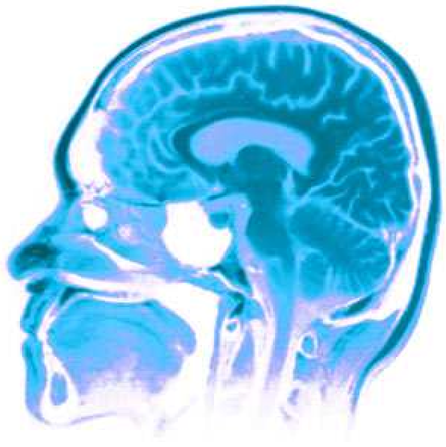




Problems with the MLE

- Computation can be expensive.
- It suffers from overfitting.
- It provides a point-estimate (best guess) of the parameters. I does not give any measure of uncertainty in this guess.
- It migh not be the best option when the data generating mechanism is outside the model class. Even if it finds the closest solution in terms of KL divergence, other distance metrics might be more appropriate.

Next class



Unconstrained optimization

Nando de Freitas

2011

KPM Book Sections: *11.2, 11.3. Please revise linear algebra in the appendix.*