Image Searching and Modelling

Part IB Paper 8
Information Engineering Elective

Lecture 2: Maximum Likelihood, Optimization, and Outlier Removal

Zoubin Ghahramani
zoubin@eng.cam.ac.uk

Department of Engineering
University of Cambridge

Easter Term

Outline

- How do we find ML parameters
  - Solving for optimum (when possible)
  - Optimization and gradient ascent
  - [Newton-Raphson optimization]
- Novelty detection and outlier removal
- Bayesian learning: example binary data and the Beta distribution

How do we find maximum likelihood (ML) parameters?

$$\boldsymbol{\theta}_{\mathrm{ML}} = \mathrm{argmax}_{\boldsymbol{\theta}}\, p(\mathcal{D}|\boldsymbol{\theta}) = \mathrm{argmax}_{\boldsymbol{\theta}} \ln p(\mathcal{D}|\boldsymbol{\theta})$$

Let's define

$$\mathcal{L}(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} \ln p(\mathcal{D}|\boldsymbol{\theta})$$

This is simply a function of $\boldsymbol{\theta}$.
At a local optimum (maxiumum):

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0, \quad \text{and} \quad \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta^2} < 0.$$

In multiple dimensions:

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0, \quad \text{and} \quad H = \frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}$$

where the *Hessian* H is negative definite.

$$H_{ij} \stackrel{\mathrm{def}}{=} \frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}$$

Aside: positive definite and negative definite matrices

A symmetric matrix M is positive definite:

- iff $\forall \mathbf{x} \neq 0$, $\mathbf{x}^\top M \mathbf{x} > 0$

- iff all eigenvalues of M are > 0

For negative definite: replace $>$ with $<$.

Solving for the optimum

Solve $\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$ and check that it's a maximum.

Multivariate Bernoulli Case:

Let $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_N\}$ be a data set of images and $\mathbf{x}_n = (x_{n1}, x_{n2}, \ldots, x_{nD})$ denote D binary features of the image, with $x_{nd} \in \{0, 1\}$.

$$P(\mathbf{x}_n|\boldsymbol{\theta}) = \prod_{d=1}^{D} \theta_d^{x_{nd}} (1 - \theta_d)^{(1-x_{nd})}$$

$$P(\mathcal{D}|\boldsymbol{\theta}) = \prod_{n=1}^{N} P(\mathbf{x}_n|\boldsymbol{\theta}) = \prod_n \prod_d \theta_d^{x_{nd}} (1 - \theta_d)^{(1-x_{nd})}$$

$$\mathcal{L}(\boldsymbol{\theta}) = \ln P(\mathcal{D}|\boldsymbol{\theta}) = \sum_{nd} x_{nd} \ln \theta_d + (1 - x_{nd}) \ln(1 - \theta_d)$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_d} = \frac{\sum_n x_{nd}}{\theta_d} - \frac{\sum_n (1 - x_{nd})}{1 - \theta_d} = 0 \quad \Rightarrow \quad \boxed{\theta_d = \frac{\sum_n x_{nd}}{N}}$$

This is very intuitive: ML parameter estimate is the frequency of 1s.

Example

$$\mathcal{D} = \begin{matrix} \{(1 & 1 & 0) \\ (1 & 0 & 0) \\ (0 & 1 & 0)\} \end{matrix}$$

$\Rightarrow$

$$\boldsymbol{\theta}_{\mathrm{ML}} = \begin{pmatrix} \frac{2}{3} & \frac{2}{3} & 0 \end{pmatrix}$$

We can similarly solve for the maximum likelihood parameters of a Gaussian.

What do we do when we can't analytically solve $\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$

Iterative optimization algorithms

- Gradient ascent (steepest ascent)

- Newton's method

Gradient ascent (steepest ascent)

Input: initial $\boldsymbol{\theta}^{(0)}$, stepsize η, convergence tolerence ϵ

Repeat:

$$\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t)} + \eta \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}$$

Until: $\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(t)}) < \epsilon$

Aside: Newton's Method

Use Taylor expansion to get local quadratic approximation to $\mathcal{L}(\theta)$:

$$\mathcal{L}(\theta) \approx \mathcal{L}(\theta^{(t)}) + (\theta - \theta^{(t)}) \frac{\partial \mathcal{L}(\theta)}{\partial \theta}|_{\theta^{(t)}} + \frac{1}{2}(\theta - \theta^{(t)})^2 \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta^2}|_{\theta^{(t)}}$$

Find maximum of this quadratic function by taking derivs wrt θ:

$$\frac{\partial \mathcal{L}}{\partial \theta} + (\theta - \theta^{(t)}) \frac{\partial^2 \mathcal{L}}{\partial \theta^2} = 0$$

$$\theta = \theta^{(t)} - \left(\frac{\partial^2 \mathcal{L}}{\partial \theta^2} \right)^{-1} \left(\frac{\partial \mathcal{L}}{\partial \theta} \right)$$

Newton's Method

Input: initial $\theta^{(0)}$, convergence tolerance ϵ

Repeat:

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \left(\left. \frac{\partial^2 \mathcal{L}}{\partial \theta^2} \right|_{\theta^{(t)}} \right)^{-1} \left(\left. \frac{\partial \mathcal{L}}{\partial \theta} \right|_{\theta^{(t)}} \right)$$

Until: $|\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(t)})| < \epsilon$

Outlier removal and novelty detection

Which is an outlier?

$\mathbf{x}_1$	1	1	0	1	0	1	1
$\mathbf{x}_2$	1	1	0	0	1	1	1
$\mathbf{x}_3$	1	0	1	0	0	1	0
$\mathbf{x}_4$	1	1	0	1	1	0	1
$\mathbf{x}_5$	1	1	0	1	1	1	1
$\boldsymbol{\theta}_{\mathrm{ML}}$	1	0.8	0.2	0.6	0.6	0.8	0.8

$$P(\mathbf{x}_3|\boldsymbol{\theta}_{\mathrm{ML}}) \approx 0.001$$

$$P(\mathbf{x}_4|\boldsymbol{\theta}_{\mathrm{ML}}) \approx 0.037$$