Inf2b Learning and Data

Lecture 11: Review: Gaussians and Linear discriminants

Hiroshi Shimodaira
(Credit: Iain Murray and Steve Renals)

Centre for Speech Technology Research (CSTR)
School of Informatics
University of Edinburgh

Jan-Mar 2014

Today's Schedule

1. Gaussian distributions
2. Maximum likelihood estimation
3. Covariance matrices

Warning: a lot of maths!

Symbol $^{(\dagger)}$: extra topics

Gaussian distributions and discriminant functions

- Univariate Gaussian pdf:

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

- Discriminant function for a univariate Gaussian pdf:

$$y_c(x) = \ln p(\mu_c, \sigma_c^2|x) = -\frac{1}{2}\frac{(x-\mu_c)^2}{\sigma_c^2} - \frac{1}{2}\ln\sigma_c^2 + \ln P(c)$$

- Multivariate Gaussian pdf:

$$p(\mathbf{x}\,|\,\boldsymbol{\mu},\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

- Discriminant function for a multivariate Gaussian pdf:

$$y_c(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_c)^T\Sigma_c^{-1}(\mathbf{x}-\boldsymbol{\mu}_c) - \frac{1}{2}\ln|\Sigma_c| + \ln P(c)$$

$$= -\frac{1}{2}\mathbf{x}^T\Sigma_c^{-1}\mathbf{x} + \boldsymbol{\mu}_c^T\Sigma_c^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_c^T\Sigma_c^{-1}\boldsymbol{\mu}_c - \frac{1}{2}\ln|\Sigma_c| + \ln P(c)$$

Parameter estimation of Gaussian distributions

- Given an observation (training) set of N samples: $D = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}\}$, where $\mathbf{x}^{(n)} \in \mathcal{R}^d$, which came from a large population.
- How can we estimate the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ of the population?
- Maximum Likelihood (ML) estimation

$$\max_{\boldsymbol{\mu},\Sigma} p(D|\boldsymbol{\mu},\Sigma)$$

- Maximum Posterior Probability (MAP) estimation $^{(\dagger)}$

$$\max_{\boldsymbol{\mu},\Sigma} p(\boldsymbol{\mu},\Sigma\,|\,D) = \max_{\boldsymbol{\mu},\Sigma} p(D\,|\,\boldsymbol{\mu},\Sigma)\; p(\boldsymbol{\mu},\Sigma)$$

ML estimation of a univariate Gaussian pdf

Assumption:

Samples $D = \{x^{(n)}\}_{n=1}^N$ were drawn independently (i.i.d)

Likelihood:

$$p(D\,|\,\mu,\sigma^2) = p(x^{(n)},\ldots,x^{(N)}\,|\,\mu,\sigma^2)$$

$$= p(x^{(1)}\,|\,\mu,\sigma^2)\cdots p(x^{(N)}\,|\,\mu,\sigma^2) = \prod_{n=1}^N p(x^{(n)}\,|\,\mu,\sigma^2)$$

$$\triangleq L(\mu,\sigma^2\,|\,D)$$

Optimisation problem:

Find the parameters μ and σ^2 that maximise the likelihood:

$$\max_{\mu,\sigma^2} L(\mu,\sigma^2\,|\,D)$$

ML estimation of a univariate Gaussian pdf

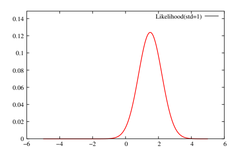
$$LL(\mu,\sigma^2\,|\,D) = \log L(\mu,\sigma^2\,|\,D) = \log\prod_{n=1}^N p(x^{(n)}\,|\,\mu,\sigma^2)$$

$$= \sum_{n=1}^N \log p(x^{(n)}\,|\,\mu,\sigma^2)$$

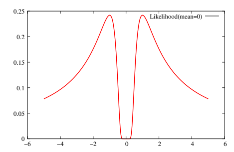
$$= \sum_{n=1}^N \log\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

$$= -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \sum_{n=1}^N\frac{(x^{(n)}-\mu)^2}{2\sigma^2}$$

Examples of likelihood function



likelihood as a function of μ



likelihood as a function of σ

$$L(\mu,\sigma^2;x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

ML estimation of a univariate Gaussian pdf

$$LL(\mu,\sigma^2\,|\,D) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \sum_{n=1}^N\frac{(x^{(n)}-\mu)^2}{2\sigma^2}$$

$$\frac{\partial LL(\mu,\sigma^2\,|\,D)}{\partial\mu} = 2\sum_{n=1}^N\frac{x^{(n)}-\mu}{2\sigma^2} = 0$$

$$\Rightarrow\ \hat{\mu} = \frac{1}{N}\sum_{n=1}^N x^{(n)}$$

$$\frac{\partial LL(\hat{\mu},\sigma^2\,|\,D)}{\partial\sigma^2} = -\frac{N}{2}\frac{1}{\sigma^2} + \sum_{n=1}^N\frac{(x^{(n)}-\hat{\mu})^2}{2(\sigma^2)^2} = 0$$

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

Point estimation vs interval estimation $^{(\dagger)}$

Estimating $p(\mathbf{x}\,|\,C)$ from training data set D, i.e. $p(\mathbf{x}\,|\,D)$

- based on point estimation (e.g. ML, MAP)

$$p(\mathbf{x}\,|\,D) = p(\mathbf{x}\,|\,\Lambda^*),\quad \Lambda^* = (\boldsymbol{\mu}^*, \sigma^{2*})$$

- based on interval estimation (Bayesian estimation)

$$p(\mathbf{x}\,|\,D) = \int p(\mathbf{x}\,|\,\Lambda, D)p(\Lambda|D)d\Lambda$$

$$= \int p(\mathbf{x}\,|\,\Lambda)p(\Lambda|D)d\Lambda$$

where $p(\Lambda\,|\,D) = \dfrac{p(D\,|\,\Lambda)p(\Lambda)}{\int p(D\,|\,\Lambda)p(\Lambda)d\Lambda}$

Covariance matrix

Sample covariance matrix:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \cdots & \sigma_{dd} \end{pmatrix} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}^{(n)} - \boldsymbol{\mu})(\mathbf{x}^{(n)} - \boldsymbol{\mu})^T$$

- Symmetric : $\boldsymbol{\Sigma}^T = \boldsymbol{\Sigma}$, and $(\boldsymbol{\Sigma}^{-1})^T = \boldsymbol{\Sigma}^{-1}$
- Positive definite: $\mathbf{x}^T\boldsymbol{\Sigma}\mathbf{x} \geq 0$, and $\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} \geq 0$

Properties of covariance matrix (†)

$$\boldsymbol{\Sigma} = V\,D\,V^T$$

$$= \begin{pmatrix} v_{11} & \cdots & v_{1d} \\ \vdots & \ddots & \vdots \\ v_{d1} & \cdots & v_{dd} \end{pmatrix} \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{pmatrix} \begin{pmatrix} v_{11} & \cdots & v_{1d} \\ \vdots & \ddots & \vdots \\ v_{d1} & \cdots & v_{dd} \end{pmatrix}^T$$

$$= (\mathbf{v}_1, \ldots, \mathbf{v}_d)\ \mathrm{Diag}(\lambda_1, \ldots, \lambda_d)\ (\mathbf{v}_1, \ldots, \mathbf{v}_d)^T$$

- $\mathbf{v}_i$: eigen vector, λ_i : eigen value
  $$\boldsymbol{\Sigma}\,\mathbf{v}_i = \lambda_i\,\mathbf{v}_i$$
- $\lambda_i \geq 0, \quad ||\mathbf{v}_i|| = 1$
- $|\boldsymbol{\Sigma}| = \prod_{i=1}^{d} \lambda_i$
- $\sum_{i=1}^{d} \sigma_{ii} = \sum_{i=1}^{d} \lambda_i$

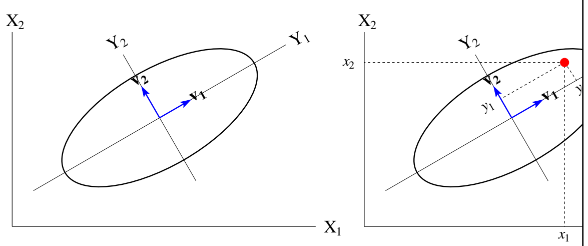
Properties of covariance matrix (†)

- $\mathrm{rank}(\boldsymbol{\Sigma})$
  - the number of linearly independent columns (or rows)
  - the number of bases (i.e. the dimension of the column space)

$$\mathrm{rank}(\boldsymbol{\Sigma}) = d \quad\rightarrow\quad \forall_i \;:\; \lambda_i > 0$$
$$\forall_{i\neq j} \;:\; \mathbf{v}_i \perp \mathbf{v}_j$$
$$|\boldsymbol{\Sigma}| > 0$$

$$\mathrm{rank}(\boldsymbol{\Sigma}) < d \quad\rightarrow\quad \exists_i \;:\; \lambda_i = 0$$
$$\exists_{(i,j)} \;:\; \mathbf{v}_i \parallel \mathbf{v}_j$$
$$|\boldsymbol{\Sigma}| = 0$$

Geometry of covariance matrix (†)



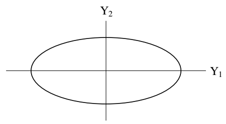
Sort eigen values: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$

$\mathbf{v}_1$: eigen vector of λ_1

$\mathbf{v}_2$: eigen vector of λ_2

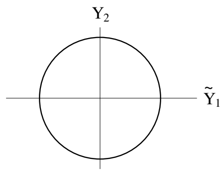
$$y_1 = \mathbf{v}_1^T\mathbf{x}\,, \quad \mathrm{Var}(y_1) = \lambda_1$$
$$y_2 = \mathbf{v}_2^T\mathbf{x}\,, \quad \mathrm{Var}(y_2) = \lambda_2$$

Geometry of covariance matrix (†)



$\longrightarrow$

$$\tilde{y}_1 = y_1/\sqrt{\lambda_1}$$
$$\tilde{y}_2 = y_2/\sqrt{\lambda_1}$$



$$(\mathbf{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \;=\; (\tilde{\mathbf{y}} - \tilde{\mathbf{u}})^T(\tilde{\mathbf{y}} - \tilde{\mathbf{u}}) \;=\; ||\tilde{\mathbf{y}} - \tilde{\mathbf{u}}||^2$$

$$\text{where } \tilde{\mathbf{u}} = \left(\frac{\mathbf{v}_1}{\sqrt{\lambda_1}}, \frac{\mathbf{v}_2}{\sqrt{\lambda_2}}\right)^T\boldsymbol{\mu}$$

Problems with the estimation of covariance matrix

- $|\boldsymbol{\Sigma}| \rightarrow 0$ when
  - the amount of training data is small
  - the dimensionality of feature vector is high
- $\boldsymbol{\Sigma}^{-1}$ gets rather unstable even if it exists
- Solutions?
- Assume a diagonal covariance matrix rather than a 'full' covariance matrix.
- Reduce the dimensionality by transforming the data into a low-dimensional vector space (PCA).
- Another regularisation:
  - Add a small positive number to the diagonal elements
    $$\boldsymbol{\Sigma} \leftarrow \boldsymbol{\Sigma} + \epsilon I$$

What if $|\Sigma_i|$ are the same for all classes? (†)

Summary

- Maximum likelihood estimation (MLE)
- Properties of covariance matrix
- Practical problem with covariance matrix estimation