The Gaussian Distribution
Machine Learning and Pattern Recognition

Chris Williams
School of Informatics, University of Edinburgh
August 2014

(All of the slides in this course have been adapted from previous versions by Charles Sutton, Amos Storkey, David Barber.)

Outline

▶ A useful model for real-valued quantities
▶ Univariate Gaussian
▶ Multivariate Gaussian
▶ Maximum likelihood estimation
▶ Class conditional classification
▶ Reading: Murphy 4.1.2, 4.1.3 (without proof), 4.2 up to end of 4.2.1; or Barber 8.4 up to start of 8.4.1 and 8.8 up to start of 8.8.2.

The Gaussian Distribution

▶ The Gaussian distribution is one of the most common distributions over continuous variables.
▶ The one dimensional Gaussian distribution is given by

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

▶ \( x \sim N(\mu, \sigma^2) \) (\( x \) is distributed as...).
▶ \( \mu \) is the mean of the Gaussian and \( \sigma^2 \) is the variance.
▶ If \( \mu = 0 \) and \( \sigma^2 = 1 \) then \( N(x; \mu, \sigma^2) \) is called a standard Gaussian.

Plot

▶ This is a standard one dimensional Gaussian distribution.
▶ All Gaussians have the same shape subject to scaling and displacement.
▶ If \( x \) is distributed \( N(\mu, \sigma^2) \), then \( y = (x - \mu)/\sigma \) is distributed \( N(0, 1) \).
Normalization

- Remember all distributions must integrate to one. The $\sqrt{\frac{2}{\pi} \sigma^2}$ is called a normalization constant - it ensures this is the case.
- Hence tighter Gaussians have higher peaks:

Maximum Likelihood Estimation

- Maximum likelihood: Set $\gamma = 1/\sigma^2$ Take derivatives
  \[
  \log P(X|\mu, \gamma) = -\frac{1}{2} \sum_{n} \gamma (x^n - \mu)^2 - \frac{N}{2} \log(2\pi) + \frac{N}{2} \log \gamma
  \]
  \[
  \frac{\partial \log P(X|\mu, \gamma)}{\partial \mu} = \gamma \sum_{n} (x^n - \mu)
  \]
  \[
  \frac{\partial \log P(X|\mu, \gamma)}{\partial \gamma} = -\frac{1}{2} \sum_{n} (x^n - \mu)^2 + \frac{N}{2\gamma}
  \]
- Hence equating derivatives to zero: $\hat{\mu} = (1/N) \sum_n x^n$ and $\hat{\sigma}^2 = (1/N) \sum_n (x^n - \hat{\mu})^2$.

Multivariate Gaussian

- The vector $x$ is multivariate Gaussian if for mean $\mu$ and covariance matrix $\Sigma$, it is distributed according to
  \[
  P(x|\mu, \Sigma) = \frac{1}{|(2\pi\Sigma)^{1/2}|} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)
  \]
- The univariate Gaussian is a special case of this.
- Shorthand: $x \sim N(\mu, \Sigma)$
- $\Sigma$ is called a covariance matrix, i.e., each element says $\sigma_{ij} = \text{Cov}(X_i, X_j)$, where
  \[
  \text{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]
  \]
- $\Sigma$ must be symmetric and positive definite

Multivariate Gaussian: Picture
Mahalanobis Distance

\[ d^2_{\Sigma}(x_i, x_j) = (x_i - x_j)^T \Sigma^{-1}(x_i - x_j) \]

- \( d^2_{\Sigma}(x_i, x_j) \) is called the Mahalanobis distance between \( x_i \) and \( x_j \)
- If \( \Sigma \) is diagonal, the contours of \( d^2_{\Sigma} \) are axis-aligned ellipsoids
- If \( \Sigma \) is not diagonal, the contours of \( d^2_{\Sigma} \) are rotated ellipsoids
  \[ \Sigma = U \Lambda U^T \]
  where \( \Lambda \) is diagonal and \( U \) is a rotation matrix (eigendecomposition of \( \Sigma \))
- \( \Sigma \) is positive definite \( \Rightarrow \) entries in \( \Lambda \) are positive

Example

- The data.

Multivariate Gaussian: Maximum Likelihood

- The Maximum Likelihood estimate can be found in the same way.
  - \( \mu = (1/N) \sum_{n=1}^{N} x^n \)
  - \( \Sigma = (1/N) \sum_{n=1}^{N} (x^n - \mu)(x^n - \mu)^T \)
- Sometimes the Gaussian is parameterized in terms of the precision matrix \( \Lambda = \Sigma^{-1} \).
Class conditional classification

Example
Suppose you have variables ‘position’ and ‘class’ where the position is a location in \( D \)-dimensional space. Suppose you have data \( D \) consisting of examples of position and class. If we assume that all the points with a particular class label are Gaussian, describe how, using the data, you could predict the class for a previously unseen position (and give the accuracy of the prediction).

Key Facts About Gaussians

- Sums of Gaussian RVs are Gaussian
- Linear Gaussian models are jointly Gaussian. In general, let
  \[
  p(x) = N(x|\mu_x, \Sigma_x)
  \]
  \[
  p(y|x) = N(y|Ax + b, \Sigma_y)
  \]
  Then \( p(x, y) \) is Gaussian, and so is \( p(x|y) \). See Murphy 4.3.
- If \( p(x, y) \) a multivariate Gaussian, both the marginals \( p(x), p(y) \) and the conditionals \( p(x|y), p(y|x) \) are Gaussian.

Inference in Gaussian models

- Partition variables into two groups, \( X_1 \) and \( X_2 \)

\[
\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}
\]

\[
\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}
\]

\[
\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)
\]

\[
\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}
\]

- For proof see e.g. §4.3.4 of Murphy (2012) (not examinable)

Class conditional classification

Learning: Fit Gaussian to data in each class (class conditional fitting). Gives \( p(position|class) \)

Find estimate for probability of each class (see last lecture) \( p(class) \)

Inference: Given a new position, we can ask “What is the probability of this point being generated by each of the Gaussians?”

Better still give probability using Bayes rule

\[
P(class|position) \propto P(position|class)P(class)
\]

Then can get ratio

\[
P(class = 1|position)/P(class = 0|position).
\]

Decision boundary for two classes is where this ratio is one.
Summary

- A useful model for real-valued quantities
- Univariate Gaussian
- Multivariate Gaussian
- Maximum likelihood estimation
- Class conditional classification