Learning from Data: Density Estimation - Gaussian Distribution

Amos Storkey, School of Informatics

Semester 1

Worked Example

- 3 class data: e.g. 1,2,1,3,1,1,2,1,2,3,1,1,1.
- Represent in 1 of m: $c_i = 1$ iff (if and only if) class is i.
- Write out likelihood of one datum. Use $\theta_i = P(c_i = 1)$

$$P(c_i = 1|\Theta) = \prod_i \theta_i^{c_i} = \theta_1^{c_1}\theta_2^{c_2}\theta_3^{c_3}$$

- (Note: $\theta_1^{c_1}\theta_2^{c_2}\theta_3^{c_3} = \theta_2$ iff $c_2 = 1$) etc.
- Now for all the data D:

$$P(D|\Theta) = \prod_\mu \prod_i \theta_i^{c_i^\mu} = \theta_1^{N_1}\theta_2^{N_2}\theta_3^{N_3}$$

- Take logs

$$\log P(D|\Theta) = \sum_\mu \sum_i c_i^\mu \log\theta_i = \sum_i N_i \log\theta_i$$

Worked Example

- Take logs

$$\log P(D|\Theta) = \sum_{\mu} \sum_{i} c_i^{\mu} \log \theta_i = \sum_{i} N_i \log \theta_i$$

- Need to optimise subject to condition $\sum_i \theta_i = 1$. Add on Lagrange multiplier term $\lambda(\sum_i \theta_i - 1)$, and differentiate wrt θ_k using

$$\frac{\partial}{\partial \theta_k} [\log P(D|\Theta) + \lambda(\sum_i \theta_i - 1)] = \frac{N_k}{\theta_k} + \lambda$$

- Set derivative to zero to get $\theta_k = -N_k/\lambda$. Substitute into constraint: $\sum_i \theta_i = 1$ to get $\lambda = -\sum_k N_k$. Final answer:

$$\theta_k = \frac{N_k}{\sum_k N_k}$$

The Gaussian Distribution

- This lecture we will be focusing on continuous quantities.
- The most common (and most easily analysed) distribution for continuous quantities is the Gaussian distribution.
- Gaussian distribution is often a reasonable model for many quantities due to various central limit theorems.
- Gaussian is sometimes called a normal distribution.

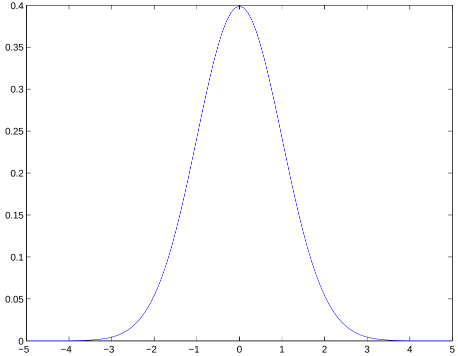
Definition

- The one dimensional Gaussian distribution is given by

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

- μ is the *mean* of the Gaussian and σ^2 is the *variance*.
- If $\mu = 0$ and $\sigma^2 = 1$ then $N(x;\mu,\sigma^2)$ is called a *standard* Gaussian.

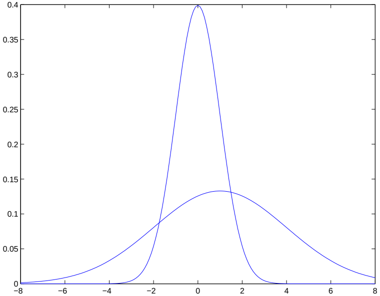
Plot



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

Normalisation

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



Central Limit Theorems (Interest Only)

- X_i mean 0, variance Σ, not necessarily Gaussian.
- X_i subject to various conditions (e.g. IID).

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i \sim N(0, \Sigma)$$

asymptotically as $N \rightarrow \infty$.

Maximum Likelihood Estimation

- Suppose we have data $\{x_i, i = 1, 2, \ldots, n\}$.
- Suppose we presume the data was generated from a Gaussian with mean μ and variance σ^2. Call this the model.
- Then the log probability of the data given the model is

$$\log \prod_i P(x_i|\mu, \sigma^2) = -\frac{1}{2}\sum_i \frac{(x_i - \mu)^2}{\sigma^2} - \frac{N}{2}\log(2\pi\sigma^2)$$

Steps left as exercise: hint $\log \prod = \sum \log$

Maximum Likelihood Estimation

- Maximum likelihood: Set $\gamma = 1/\sigma^2$ Take derivatives

$$\log P(X|\mu,\gamma) = -\frac{1}{2}\sum_i \gamma(x_i-\mu)^2 - \frac{N}{2}\log(2\pi) + \frac{N}{2}\log\gamma$$

$$\frac{\partial \log P(X|\mu,\gamma)}{\partial\mu} = \gamma\sum_i (x_i-\mu)$$

$$\frac{\partial \log P(X|\mu,\gamma)}{\partial\gamma} = -\frac{1}{2}\sum_i (x_i-\mu)^2 + \frac{N}{2\gamma}$$

- Hence $\mu = (1/N)\sum_i x_i$ and $\sigma^2 = (1/N)\sum_i (x_i-\mu)^2$.
- (Maximum likelihood estimate of σ^2 is *biased*.)

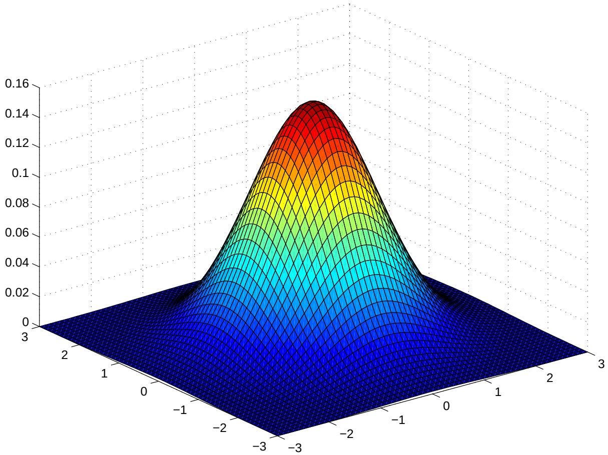
Multivariate Gaussian

- The vector $\mathbf{x}$ is multivariate Gaussian if for mean $\boldsymbol{\mu}$ and covariance matrix Σ, it is distributed according to

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

- The univariate Gaussian is a special case of this.
- Σ is called a covariance matrix. It says how much attributes co-vary. More later.

Multivariate Gaussian: Picture

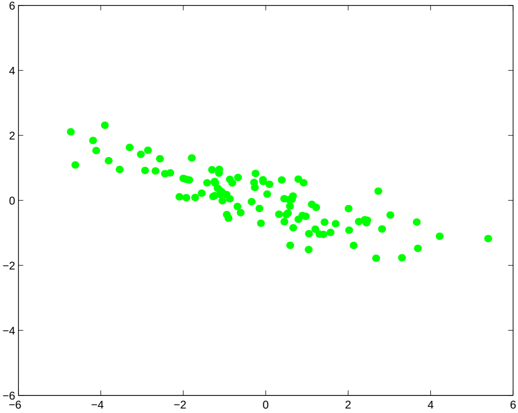


Multivariate Gaussian: Maximum Likelihood

- The Maximum Likelihood estimate can be found in the same way.
- $\boldsymbol{\mu} = (1/N)\sum_{i=1}^{N} \mathbf{x}_i$
- $\Sigma = (1/N)\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T$

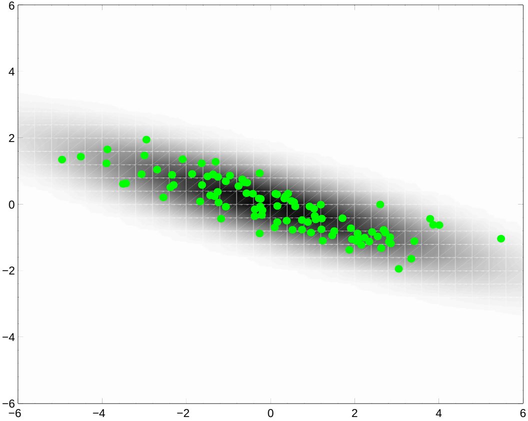
Example

- The data.



Example

- The data. The maximum likelihood fit.



Class conditional classification

- Have real valued multivariate data, along with class label for each point.
- Want to predict the value of the class label given some new point.
- Presume that if we take all the points with a particular label, then we believe they were sampled from a Gaussian.
- How should we predict the class at a new point?

Class conditional classification

- Learning: Fit Gaussian to data in each class (class conditional fitting). Gives $P(\text{position}|\text{class})$
- Find estimate for probability of each class (see last lecture) $P(\text{class})$
- Inference: Given a new position, we can ask "What is the probability of this point being generated by each of the Gaussians?"
- Pick the largest (just like maximum likelihood)
- Better still give probability using Bayes rule

  $$P(\text{class}|\text{position}) \propto P(\text{position}|\text{class})P(\text{class})$$

  Then can get ratio
  $P(\text{class} = 1|\text{position})/P(\text{class} = 0|\text{position})$.
- Decision boundary for two classes is where this ratio is one.

Summary

- Gaussian
- Maximum Likelihood fitting of a Gaussian
- Multivariate Gaussian and covariances again.
- Maximum Likelihood fitting.
- Class conditional classification using Gaussians.