UNIVERSITY
of
GLASGOW

Machine Learning

Lecture. 8.

Mark Girolami

girolami@dcs.gla.ac.uk

Department of Computing Science
University of Glasgow

Classification

UNIVERSITY
of
GLASGOW

- The previous approach to classification focused on modeling the discriminant function directly using a linear model i.e.

$$\log \frac{P(C=1|\mathbf{x})}{P(C=0|\mathbf{x})} = \mathbf{w}^{\mathsf{T}}\phi(\mathbf{x})$$

Classification

UNIVERSITY *of* GLASGOW

- The previous approach to classification focused on modeling the discriminant function directly using a linear model i.e.

$$\log \frac{P(C=1|\mathbf{x})}{P(C=0|\mathbf{x})} = \mathbf{w}^{\mathsf{T}}\phi(\mathbf{x})$$

- The generative approach on the other hand seeks to define the discriminant function by directly estimating the posterior ratio from the data likelihood and prior terms i.e.

$$\log \frac{P(C=1|\mathbf{x})}{P(C=0|\mathbf{x})} = \log \frac{P(\mathbf{x}|C=1)P(C=1)}{P(\mathbf{x}|C=0)P(C=0)}$$

Classification

UNIVERSITY
of
GLASGOW

- Now given a training data set, $\mathbf{X}, \mathbf{t}$, we can estimate the prior probabilities of class membership by simply counting the numbers of instance of each class in the data and normalising by the total number of data samples i.e.

$$\widehat{P}(C = k) = \frac{1}{N_k} \sum_{n=1}^{N} \delta(t_n, k)$$

Classification

UNIVERSITY *of* GLASGOW

- Now given a training data set, $\mathbf{X}, \mathbf{t}$, we can estimate the prior probabilities of class membership by simply counting the numbers of instance of each class in the data and normalising by the total number of data samples i.e.

$$\widehat{P}(C = k) = \frac{1}{N_k} \sum_{n=1}^{N} \delta(t_n, k)$$

- Note that the *hat* notation is being used to indicate that we are estimating the probability of class membership from this finite data sample.

Classification

UNIVERSITY
of
GLASGOW

- Now we require the class conditional data-likelihood $P(\mathbf{x}|C = k)$, that is the probability density or distribution from which the data is generated.

Classification

UNIVERSITY
of
GLASGOW

- Now we require the class conditional data-likelihood $P(\mathbf{x}|C = k)$, that is the probability density or distribution from which the data is generated.
- We will look at this important and general problem, probability density estimation, in the first two lectures devoted to Unsupervised Learning.

Classification

UNIVERSITY
of
GLASGOW

- Now we require the class conditional data-likelihood $P(\mathbf{x}|C = k)$, that is the probability density or distribution from which the data is generated.
- We will look at this important and general problem, probability density estimation, in the first two lectures devoted to Unsupervised Learning.
- However, for now we will look at two specific situations where we can make assumptions about the parametric form of the class-conditional likelihoods.

Classification

UNIVERSITY
of
GLASGOW

- Let us for now assume that we have reason to believe that our class-conditional likelihoods are well represented by multivariate Gaussians such that

$$p(\mathbf{x}|C=k) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}_k|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^{\mathsf{T}}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right.$$

Classification

UNIVERSITY
of
GLASGOW

- Let us for now assume that we have reason to believe that our class-conditional likelihoods are well represented by multivariate Gaussians such that

$$p(\mathbf{x}|C=k) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}_k|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^{\mathsf{T}}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right.$$

- Then we require to obtain estimates for the mean vectors $\widehat{\boldsymbol{\mu}}_k$ and the covariance matrix $\widehat{\boldsymbol{\Sigma}}_k$ to obtain our estimated class-conditional likelihood $\widehat{p}(\mathbf{x}|C=k)$ which can be plugged into our discriminant function.

Classification

- Lets expand the discriminant function for two classes, say k and l then it is easy to show that

$$\begin{aligned}\log \frac{P(C=k|\mathbf{x})}{P(C=l|\mathbf{x})} &= \log \frac{P(\mathbf{x}|C=k)}{P(\mathbf{x}|C=l)} + \log \frac{P(C=k)}{P(C=l)} \\ &= \frac{1}{2}\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} + \mathbf{w}^{\mathsf{T}}\mathbf{x} + b_0\end{aligned}$$

Classification

UNIVERSITY of GLASGOW

- Lets expand the discriminant function for two classes, say k and l then it is easy to show that

$$\begin{aligned}\log \frac{P(C=k|\mathbf{x})}{P(C=l|\mathbf{x})} &= \log \frac{P(\mathbf{x}|C=k)}{P(\mathbf{x}|C=l)} + \log \frac{P(C=k)}{P(C=l)} \\ &= \frac{1}{2}\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} + \mathbf{w}^{\mathsf{T}}\mathbf{x} + b_0\end{aligned}$$

- where $\mathbf{A} = \boldsymbol{\Sigma}_l^{-1} - \boldsymbol{\Sigma}_k^{-1}$ and $\mathbf{w} = \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \boldsymbol{\Sigma}_l^{-1}\boldsymbol{\mu}_l$ with

$$b_0 = \log \frac{P(C=k)}{P(C=l)} + \frac{1}{2}\log \frac{|\boldsymbol{\Sigma}_l|}{|\boldsymbol{\Sigma}_k|} + \frac{1}{2}\left(\boldsymbol{\mu}_l^{\mathsf{T}}\boldsymbol{\Sigma}_l^{-1}\boldsymbol{\mu}_l - \boldsymbol{\mu}_k^{\mathsf{T}}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\right)$$

Classification

UNIVERSITY
of
GLASGOW

- So what we can see is that the discriminant function that we obtain when assuming multivariate Gaussian class-conditional densities is a quadratic function of the features $\mathbf{x}$ and so we have a quadratic decision surface.

Classification

UNIVERSITY
of
GLASGOW

- So what we can see is that the discriminant function that we obtain when assuming multivariate Gaussian class-conditional densities is a quadratic function of the features $\mathbf{x}$ and so we have a quadratic decision surface.
- It should also be clear that if a common covariance matrix across all classes is assumed then our discriminant reduces to a linear function of the form $\mathbf{w}^{\mathsf{T}}\mathbf{x} + b_0$ where $\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l)$ which relies on the difference in the class means.

Classification

UNIVERSITY
of
GLASGOW

- Now we have to estimate the parameters of the conditional-likelihood, in this case mean and covariances, to obtain the required posterior class probabilities

Classification

UNIVERSITY
of
GLASGOW

- Now we have to estimate the parameters of the conditional-likelihood, in this case mean and covariances, to obtain the required posterior class probabilities
- As we are really only interested in the discriminant function at the end of the day then it can be argued that most effort should focus on estimating a functional form for the posterior log-likelihood ratio as in the discriminative approach.

Classification

UNIVERSITY
of
GLASGOW

- Now we have to estimate the parameters of the conditional-likelihood, in this case mean and covariances, to obtain the required posterior class probabilities
- As we are really only interested in the discriminant function at the end of the day then it can be argued that most effort should focus on estimating a functional form for the posterior log-likelihood ratio as in the discriminative approach.
- The generative approach on the other hand requires to make good estimates of the density to obtain the discriminant function and this can be a weakness of the method in that requiring data from the regions of high density for each class to estimate parameter values e.g. mean values, may not necessarily help in defining the discriminant function.

Naive Bayes

UNIVERSITY
of
GLASGOW

- In **Bioinformatics** microarray data can be used to build classifiers which will be capable of discriminating between cancerous and healthy tissue samples.

Naive Bayes

UNIVERSITY
of
GLASGOW

- In **Bioinformatics** microarray data can be used to build classifiers which will be capable of discriminating between cancerous and healthy tissue samples.
- Each sample is defined by the amount of mRNA that a large numbers of gene express in healthy or diseased conditions. Often there are over 30,000 genes, so this means that we have a feature vector $\mathbf{x} \in \mathbb{R}^D$ where $D = 30,000$.

Naive Bayes

UNIVERSITY
of
GLASGOW

- In **Bioinformatics** microarray data can be used to build classifiers which will be capable of discriminating between cancerous and healthy tissue samples.
- Each sample is defined by the amount of mRNA that a large numbers of gene express in healthy or diseased conditions. Often there are over 30,000 genes, so this means that we have a feature vector $\mathbf{x} \in \mathbb{R}^D$ where $D = 30,000$.
- If we assume that the mRNA levels are roughly Gaussian then we can see that estimating $\Sigma_{healthy}$ a $30,000 \times 30,000$ dimensional covariance matrix is going to be impossible given that the number of samples will be as small as several dozen.

Naive Bayes

UNIVERSITY
of
GLASGOW

- So despite there possibly being features which will be correlated with each other it is impractical to even consider attempting to estimate a full covariance. So we are forced to make a further assumption that the covariance matrix is diagonal such that

Naive Bayes

UNIVERSITY of GLASGOW

- So despite there possibly being features which will be correlated with each other it is impractical to even consider attempting to estimate a full covariance. So we are forced to make a further assumption that the covariance matrix is diagonal such that
-

$$\mathbf{\Sigma}_k = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_3^2 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \cdots & 0 \\ 0 & \cdots & 0 & \sigma_{D-1}^2 & 0 \\ 0 & \cdots & \cdots & 0 & \sigma_D^2 \end{pmatrix}$$

Naive Bayes

UNIVERSITY
of
GLASGOW

- In this case then the multivariate Gaussian reduces to a product form such that

$$p(\mathbf{x}|C=k) = \prod_{d=1}^{D} p(x_d|C_k) = \prod_{d=1}^{D} \mathcal{N}_{x_d}(\mu_d, \sigma_d)$$

Naive Bayes

UNIVERSITY
of
GLASGOW

- In this case then the multivariate Gaussian reduces to a product form such that

$$p(\mathbf{x}|C=k) = \prod_{d=1}^{D} p(x_d|C_k) = \prod_{d=1}^{D} \mathcal{N}_{x_d}(\mu_d, \sigma_d)$$

- Despite this form of classifier being referred to as **Naive Bayes** or **Idiots Bayes**, presumably because of the naive assumption of there being no covariance between features, in many applications such a classifier works surprisingly well.

Naive Bayes

UNIVERSITY
of
GLASGOW

- In this case then the multivariate Gaussian reduces to a product form such that

$$p(\mathbf{x}|C=k) = \prod_{d=1}^{D} p(x_d|C_k) = \prod_{d=1}^{D} \mathcal{N}_{x_d}(\mu_d, \sigma_d)$$

- Despite this form of classifier being referred to as **Naive Bayes** or **Idiots Bayes**, presumably because of the naive assumption of there being no covariance between features, in many applications such a classifier works surprisingly well.
- One particular application within **Information Retrieval** is document classification which we shall look at briefly here.

Naive Bayes

UNIVERSITY
of
GLASGOW

- Assume a number of Documents d each have (or have not) the occurrence of words w from a dictionary $\mathcal{D}$.

Naive Bayes

UNIVERSITY
of
GLASGOW

- Assume a number of Documents d each have (or have not) the occurrence of words w from a dictionary $\mathcal{D}$.
- Assume a bag-of-words document model i.e $|\mathcal{D}|$ defined by single draws from a binomial distribution, so word w occurs in documents from class k with probability p_{kw} and the probability of it not occurring in the class k document is obviously $1 - p_{kw}$.

Naive Bayes

UNIVERSITY
of
GLASGOW

- Assume a number of Documents d each have (or have not) the occurrence of words w from a dictionary $\mathcal{D}$.
- Assume a bag-of-words document model i.e $|\mathcal{D}|$ defined by single draws from a binomial distribution, so word w occurs in documents from class k with probability p_{kw} and the probability of it not occurring in the class k document is obviously $1 - p_{kw}$.
- If word w occurs in the document at least once assign value 1 and if it does not occur take value 0. Each document represented by a of ones and zeros with the same length as the size of the dictionary.

Naive Bayes

UNIVERSITY
of
GLASGOW

- Clearly for large dictionaries we will need to employ a Naive Bayes classifier.

Naive Bayes

UNIVERSITY
of
GLASGOW

- Clearly for large dictionaries we will need to employ a Naive Bayes classifier.
- Define matrix $\mathbf{D}$, rows corresponding to each document and columns representing dictionary terms, so element $\mathbf{D}_{dw}$ indicates presence or absence of word w in document d.

Naive Bayes

UNIVERSITY
of
GLASGOW

- Clearly for large dictionaries we will need to employ a Naive Bayes classifier.
- Define matrix $\mathbf{D}$, rows corresponding to each document and columns representing dictionary terms, so element $\mathbf{D}_{dw}$ indicates presence or absence of word w in document d.
- Using Naive Bayes class-conditional probability of a document d coming from class k is

$$p(\mathbf{D}_d|C=k) = \prod_{w=1}^{|\mathcal{D}|} p(\mathbf{D}_{dw}|C_k) = \prod_{w=1}^{D} p_{kw}^{\mathbf{D}_{dw}}(1-p_{kw})^{1-\mathbf{D}_{dw}}$$

Naive Bayes

UNIVERSITY
of
GLASGOW

- Once the parameter values p_{kw} are estimated then the estimate of the class conditional likelihood can be plugged into the discriminant function to make classification.

Naive Bayes

UNIVERSITY
of
GLASGOW

- Once the parameter values p_{kw} are estimated then the estimate of the class conditional likelihood can be plugged into the discriminant function to make classification.
- We will see in subsequent lectures that the Maximum-Likelihood estimate for the parameters p_{kw} is simply

$$\widehat{p}_{kw} = \frac{1}{N_k} \sum_{d \in C_k} \mathbf{D}_{dw}$$

Naive Bayes

UNIVERSITY
of
GLASGOW

- Once the parameter values p_{kw} are estimated then the estimate of the class conditional likelihood can be plugged into the discriminant function to make classification.
- We will see in subsequent lectures that the Maximum-Likelihood estimate for the parameters p_{kw} is simply

$$\widehat{p}_{kw} = \frac{1}{N_k} \sum_{d \in C_k} \mathbf{D}_{dw}$$

- So if a term does not occur in the documents from class k then $\widehat{p}_{kw} = 0$ which seems a little pessimistic as it may be that additional documents from the class may well have the word.

Naive Bayes

UNIVERSITY
of
GLASGOW

- It is also somewhat inconvenient in that if $\widehat{p}_{kw} = 0$ for one word then $p(\mathbf{D}_d|C = k) = 0$ which makes no real sense.

Naive Bayes

UNIVERSITY
of
GLASGOW

- It is also somewhat inconvenient in that if $\widehat{p}_{kw} = 0$ for one word then $p(\mathbf{D}_d|C = k) = 0$ which makes no real sense.
- In further lectures we will look at Bayesian estimates of distribution parameters and we will see for binary variables that the MAP estimator is a more reasonable, and computationally convenient,

$$\widehat{p}_{kw} = \frac{1 + \sum_{d \in C_k} \mathbf{D}_{dw}}{2 + N_k}$$