Department of Mathematics and Computer Science
University of Southern Denmark, Odense

February 22, 2013

Marco Chiarandini

DM825 - Introduction to Machine Learning

Sheet 7, Spring 2013

Exercise 1 – Linear discriminants

1. Develop analytically the formulas of a generative algorithm with Gaussian likelihood for a k-way classification problem. In particular, estimate the model parameters.

2. Derive the explicit formula of the decision boundaries in the case of two predictor variables.

   Solution For any linear model used for classifying the k class we have $h_k(\vec{x}) = \vec{\theta}^T\vec{x}$. Then the boundary between two classes given by $h_k(\vec{x}) = h_l(\vec{x})$ is the set $\{\vec{x} : (\vec{\theta}_k - \vec{\theta}_l)\vec{x} = 0\}$

   Several methods use a discriminant function δ_k to map the linear model to something else. Among these methods we have those that map to a probability measure to be considered the posterior $p(\mathcal{C}_k \mid \vec{X} = \vec{x})$. For example, logistic regression uses:

$$p(\mathcal{C}_1 \mid \vec{X} = \vec{x}) = \frac{1}{1 + \exp \vec{\theta}\vec{x}}$$

$$p(\mathcal{C}_2 \mid \vec{X} = \vec{x}) = \frac{\exp \vec{\theta}\vec{x}}{1 + \exp \vec{\theta}\vec{x}}$$

   If the discrimninat function is monotone then the decision boundaries are linear. For the case of the logit transformation $\log[p/(1-p)]$ we see that

$$\log \frac{p(\mathcal{C}_1 \mid \vec{X} = \vec{x})}{p(\mathcal{C}_2 \mid \vec{X} = \vec{x})} = \vec{\theta}\vec{x}$$

   The decision boundary is the case where the two posterior probabilities are equal and hence $\log 1 = 0 = \vec{\theta}\vec{x}$.

   In the linear discriminat analysis we have:

$$\begin{aligned}
\log \frac{p(\mathcal{C}_k \mid \vec{X} = \vec{x})}{p(\mathcal{C}_l \mid \vec{X} = \vec{x})} &= \log \frac{p(\vec{x} \mid \mathcal{C}_k)p(\mathcal{C}_k)}{p(\vec{x} \mid \mathcal{C}_k)p(\mathcal{C}_k)} \\
&= \log \frac{N(\mu_k, \Sigma)\phi_k}{N(\mu_l, \Sigma)\phi_l} = \log \frac{N(\mu_k, \Sigma)}{N(\mu_l, \Sigma)} + \log \frac{\phi_k}{\phi_l} = \\
&= \log \frac{\phi_k}{\phi_l} - \frac{1}{2}(\mu_k + \mu_l)^T\Sigma^{-1}(\mu_k - \mu_l) + x^T\Sigma^{-1}(\mu_k - \mu_l)
\end{aligned}$$

   which is the linear discriminant that we were asked to find.

3. Implement the analysis in R using the data:

```
Iris <- data.frame(cbind(iris[,c(2,3)], Sp = rep(c("s","c","v"), rep
    (50,3))))
train <- sample(1:150, 75)
table(Iris$Sp[train])
```

   Plot the contour of the Gaussian distribution and linear discriminant

4. Compare your results with those of the `lda` function from the package MASS in R.

   Deepening: read section 4.3.3 of B2 and inspect the outcome of `lda` when run on the full data with all 4 predictors, ie:

```
Iris <- data.frame(cbind(iris, Sp = rep(c("s","c","v"), rep(50,3))))
z <- lda(Sp ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width
    ,
        Iris, prior = c(1,1,1)/3, subset = train)
# predict(z, Iris[-train, ])$class

plot(z,dimen=1)
plot(z,type="density",dimen=1)
plot(z,dimen=2)
```

Exercise 2 – Naive Bayes

You decide to make a text classifier. To begin with you attempt to classify documents as either sport or politics. You decide to represent each document as a (row) vector of attributes describing the presence or absence of words.

$$\vec{x} = (\text{goal, football, golf, defence, offence, wicket, office, strategy})$$

Training data from sport documents and from politics documents is represented below using a matrix in which each row represents a (row) vector of the 8 attributes.

$$\mathbf{x}_{\text{politics}} = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{x}_{\text{sport}} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 \end{bmatrix}$$

Using a Naive Bayes classifier, what is the probability that the document $\vec{x} = (1, 0, 0, 1, 1, 1, 1, 0)$ is about politics?

Solution

Step 1: Let x_j be the presence of the jth word among the 8 words. Let y be the classifiction in politics or sports. We estimate the parameters of the Bernoulli distributions involved, namely:

$$\begin{aligned} p(y) &\sim \phi_y \\ \forall j: p(x_j = 1|y = 0) &\sim \phi_{j|y=0} \\ \forall j: p(x_j = 1|y = 1) &\sim \phi_{j|y=1} \end{aligned}$$

Step 2: We do this using the joint likelihood. For a single sample among the 6+7 given, we assume x_js are conditionally independent given y. By chain rule:

$$\begin{aligned} p(x_1, \dots, x_8) &= p(x_1 \mid y)p(x_2 \mid y, x_1)p(x_3 \mid y, x_1, x_2) \dots \\ &= p(x_1 \mid y)p(x_2 \mid y)p(x_3 \mid y) \dots && \text{cond. indep.} \\ &= \prod_{i=1}^{m} p(x_i \mid y) \end{aligned}$$

$$\begin{aligned} l(\phi_y, \phi_{j|y=0}, \phi_{j|y=1}) &= \prod_{i=1}^{m} p(\vec{x}^i, y^i) \\ &= \prod_{i=1}^{m} \prod_{j=1}^{n} p(x_j^i \mid y^i) p(y^i) \end{aligned}$$

where $m = 13$ and $n = 8$. Maximizing we find: Solution:

$$\begin{aligned} \phi_y &= \frac{\sum_i I\{y^i = 1\}}{m} \\ \phi_{j|y=1} &= \frac{\sum_i I\{y^i = 1, x_j^i = 1\}}{\sum_i I\{y^i = 1\}} \\ \phi_{j|y=0} &= \frac{\sum_i I\{y^i = 0, x_j^i = 1\}}{\sum_i I\{y^i = 0\}} \end{aligned}$$

Hence it is a counting task. Working on numerical data with the help of R:

```
P <- matrix(scan("p.txt"),ncol=8,nrow=6,byrow=TRUE)
S <- matrix(scan("s.txt"),ncol=8,nrow=7,byrow=FALSE)
(phiy <- nrow(P)/(nrow(P)+nrow(S)))
0.46154
(phijy1 <- apply(P,2,sum)/nrow(P))
[1] 0.33333 0.16667 0.16667 0.83333 0.83333 0.16667 0.66667 0.83333
(phijy0 <- apply(S,2,sum)/nrow(S))
0.28571 0.14286 0.42857 0.42857 0.42857 0.28571 0.42857 0.71429
```

Step 3 and 4: To predict we maximize:

$$\begin{aligned} \arg\max_y p(y \mid \vec{x}) &= \arg\max_y \frac{p(\vec{x} \mid y)p(y)}{p(\vec{x})} \\ &\arg\max_y p(\vec{x} \mid y)p(y) \end{aligned}$$

$$p(y=1 \mid \vec{x}) = p(\vec{x} \mid y=1)p(y=1) = \prod_{j=1}^{8} p(x_j^i | y^i = 1)p(y^i = 1)$$
$$= \prod_{j=1}^{8} \phi_{j|y=1}^{x_j^i} (1-\phi_{j|y=1})^{1-x_j^i} \phi_y$$
$$p(y=0 \mid \vec{x}) = p(\vec{x} \mid y=0)p(y=0) = \prod_{j=1}^{8} p(x_j^i | y^i = 0)p(y^i = 0)$$
$$= \prod_{j=1}^{8} \phi_{j|y=0}^{x_j^i} (1-\phi_{j|y=0})^{1-x_j^i} (1-\phi_y)$$

In logarithms:

$$\log p(y=1 \mid \vec{x}) = \sum_{j=1}^{8} [x_j^i \log(\phi_{j|y=1}) + (1-x_j^i)\log(1-\phi_{j|y=1})] + \log(\phi_y)$$
$$\log p(y=0 \mid \vec{x}) = \sum_{j=1}^{8} [x_j^i \log(\phi_{j|y=0}) + (1-x_j^i)\log(1-\phi_{j|y=0})] + \log(1-\phi_y)$$

```
x <- c(1, 0, 0, 1, 1, 1, 1, 0)
py1x <- sum(log(phijy1[x]))+sum(log(1-phijy1[!x]))+log(phiy)
py0x <- sum(log(phijy0[x]))+sum(log(1-phijy0[!x]))+log(1-phiy)
py1x
-8.4227
py0x
-8.8494
```

Hence since $\log p(y=1 \mid \vec{x}) = -8.4227$ is greater than $\log p(y=0 \mid \vec{x}) = -8.8494$ we classify $\vec{x}$ as politics.