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

February 4, 2013
Marco Chiarandini

DM825 - Introduction to Machine Learning

Sheet 3, Spring 2013

Exercise 1

Redo exercise 1 from Sheet 1 using logistic regression (transform the response label -1 to 0). Alternatively use logistic regression on these data [classification.data]. Although, as we will see, logistic regression can be implemented in R via `glm`, you are asked here to implement the method by yourself. For the optimization you can reuse the gradient descent method developed in previous exercises or you can use `optim`.

Exercise 2

In exercise 3 of Sheet 2 use 1/2 of the data for training the models, 1/4 of the data to *select* the model (k-nearest neighbor or linear regression) and 1/4 to *assess* the performance of the best model selected.

Exercise 3 Bayesian prediction

In class we saw an example with binary variables. Often however we encounter discrete variables that can take on one of K possible mutually exclusive states. A way to handle this situation is to express such variables by a K-dimensional vector $\vec{x}$ in which one of the x_k elements equals to 1 and all remaining elements equal 0. Consider a sample described by m multinomial random variables $(X^1, X^2, \ldots, X^m)$, where $X^i \sim \text{Mult}(\theta)$ for each m, and where the X^i are assumed conditionally independent given θ. Let $\theta \sim \text{Dir}(\boldsymbol{\alpha})$. Now consider a random variable $X_{new} \sim \text{Mult}(\theta)$ that is assumed conditionally independent of $(X^1, X^2, \ldots, X^m)$ given θ. Compute the predictive distribution:

$$p(x_{new}|x_1, x_2, \ldots, x_N, \alpha)$$

by integrating over θ.

Solution The exercise refers to the theory developed in sec. 2.1 and 2.2 of [B1].

With multinomial distributions we consider the representation in which X_j is a random vector consisting of all 0's and a single 1. For example, $\vec{x} = (0, 0, 1, 0, 0, 0)^T$. If we denote $p(x_k = 1) = \theta_k$ then $X_j \sim \text{Mult}(\theta)$ corresponds to saying:

$$p(\vec{x}|\vec{\theta}) = \prod_{k=1}^{K} \theta_k^{x_k} \tag{1}$$

and $\vec{\theta} = (\theta_1, \ldots, \theta_K)^T$. This distribution is also known as generalized Bernoulli distribution.

Consequently, the likelihood for the training set $(X_1, X_2, \ldots, X_N)$ of independent observations is:

$$p(\vec{x}_1, \ldots, \vec{x}_N|\vec{\theta}) = \prod_{j=1}^{m}\prod_{k=1}^{K} \theta_k^{x_{jk}} = \prod_{k=1}^{K} \theta_k^{\sum_{j=1}^{m} x_{jk}} = \prod_{k=1}^{K} \theta_k^{l_k}$$

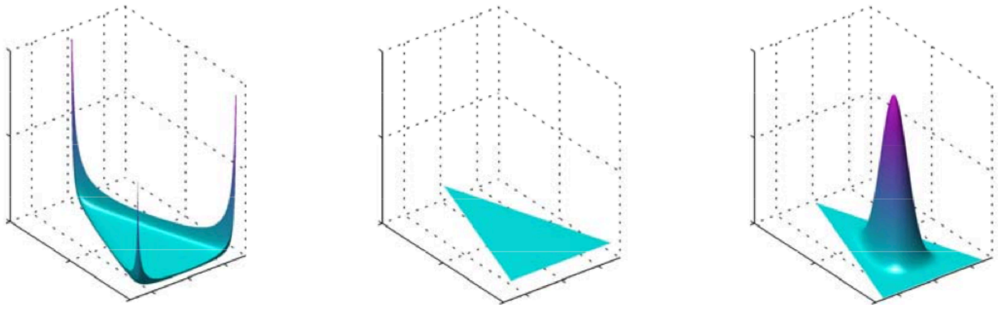

Figure 2.5 Plots of the Dirichlet distribution over three variables, where the two horizontal axes are coordinates in the plane of the simplex and the vertical axis corresponds to the value of the density. Here $\{\alpha_k\} = 0.1$ on the left plot, $\{\alpha_k\} = 1$ in the centre plot, and $\{\alpha_k\} = 10$ in the right plot.

where we let l_k be the total number of x_j that belong to class k. The prior distribution of θ is

$$\mathrm{Dir}(\vec{\theta}|\vec{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

with $0 \leq \theta_k \leq 1$, $\sum_k \theta_k = 1$, $\vec{\alpha} = (\alpha_1, \ldots, \alpha_K)^T$ and $\alpha_0 = \sum_k \alpha_k$. The Dirichlet distribution is constructed with the aim of satisfying the conjugacy property. The fraction in front of the product is the normalizing coefficient derived from:

$$\frac{1}{g(\vec{\alpha})} \int \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \mathrm{d}\vec{\theta} = 1 \tag{2}$$

$$g(\vec{\alpha}) = \frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}{\Gamma(\alpha_0)} \tag{3}$$

The expected value for the kth component of the random variable $\vec{\theta}$ is

$$\mathrm{E}[\theta_k] = \frac{\alpha_k}{\alpha_0}.$$

In the Figure an example of Dirichelt distribution in 3 dimensions and different values of α_k.
From Bayes' Theorem

$$p(\vec{\theta}|\vec{x}_1, \ldots, \vec{x}_N) \propto p(\vec{x}_1, \ldots, \vec{x}_N|\vec{\theta})p(\vec{\theta}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k + l_k - 1}$$

The posterior takes again the form of a Dirichlet distribution (conjugacy property) and comparing with the definition of the Dirichlet distribution above we can determine the normalization coefficients as

$$p(\vec{\theta}|\vec{x}_1, \ldots, \vec{x}_N) = \mathrm{Dir}(\vec{\theta}|\vec{\alpha} + \vec{l}) = \frac{\Gamma(\alpha_0 + m)}{\Gamma(\alpha_1 + l_1)\cdots\Gamma(\alpha_K + l_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k + l_k - 1} \tag{4}$$

with $\vec{l} = (l_1, \ldots, l_K)^T$.
To evaluate the predictive distribution of a new outcome we use the sum and product rules of probability

$$p(\vec{x}_{new}|\vec{x}_1, \ldots, \vec{x}_N, \alpha) = \int_{\vec{0}}^{\vec{1}} p(\vec{x}_{new}|\vec{\theta}, \vec{\alpha}) p(\vec{\theta}|\vec{x}_1, \ldots, \vec{x}_N, \vec{\alpha}) \mathrm{d}\vec{\theta}$$

From (1) and (4) we have

$$\begin{aligned}
p(\vec{x}_{new}|\vec{x}_1,\dots,\vec{x}_N,\alpha) &= \int p(\vec{x}_{new}|\vec{\theta},\vec{\alpha})p(\vec{\theta}|\vec{x}_1,\dots,\vec{x}_N,\vec{\alpha})\mathrm{d}\vec{\theta} \\
&= \int \prod_{k=1}^{K} \theta_k^{x_{new,k}} \frac{1}{g(\vec{\alpha}+\vec{l})} \prod_{k=1}^{K} \theta_k^{\alpha_k+l_k-1}\mathrm{d}\vec{\theta} \\
&= \frac{1}{g(\vec{\alpha}+\vec{l})} \int \prod_{k=1}^{K} \theta_k^{\alpha_k+l_k+x_{new,k}-1} \\
&= \frac{g(\vec{\alpha}+\vec{x}_{new}+\vec{l})}{g(\vec{\alpha}+\vec{l})} \\
p(x_{new,k}=1|\vec{x}_1,\dots,\vec{x}_N,\alpha) &= \frac{\Gamma(\alpha_1+l_1+x_1)\cdots\Gamma(\alpha_K+l_K+x_k)\Gamma(\alpha_0+m)}{\Gamma(\alpha_0+m+1)\Gamma(\alpha_1+l_1)\cdots\Gamma(\alpha_K+l_K)} \\
&= \frac{\alpha_k+l_k}{\alpha_0+m}
\end{aligned}$$

where in the last step we used the fact that $\Gamma(x+1) = x\Gamma(x)$