UNIVERSITY COLLEGE LONDON
Faculty of Engineering Sciences

Department of Computer Science

Problem Set: Classification

Dr. Dariush Hosseini (dariush.hosseini@ucl.ac.uk)

Notation

Inputs:

$\mathbf{x} = [1, x_1, x_2, ..., x_m]^T \in \mathbb{R}^{m+1}$

Outputs:

$y \in \mathbb{R}$ for regression problems

$y \in \{0, 1\}$ for binary classification problems

Training Data:

$\mathcal{S} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$

Input Training Data:

The design matrix, $\mathbf{X}$, is defined as:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}^{(1)T} \\ \mathbf{x}^{(2)T} \\ \cdot \\ \cdot \\ \mathbf{x}^{(n)T} \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & \cdot & \cdot & x_m^{(1)} \\ 1 & x_1^{(2)} & \cdot & \cdot & x_m^{(2)} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_1^{(n)} & \cdot & \cdot & x_m^{(n)} \end{bmatrix}$$

Output Training Data:

$$\mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \cdot \\ \cdot \\ y^{(n)} \end{bmatrix}$$

Data-Generating Distribution:

The outcomes of $\mathcal{S}$ are drawn i.i.d. from a data-generating distribution, $\mathcal{D}$

1. *This problem focuses on generative approaches to classification. It begins by asking for basic statements and derivations pertaining to probabilistic classification, before asking you to consider a particular generative model. The model is not one which we discussed in lectures, but is very similar to Naive Bayes. It is known as 'Linear Discriminant Analysis' (LDA). You are asked to investigate the discriminant boundaries that emerge from this model. Following this you are asked to consider a slight generalisation of the model with fewer restrictions placed upon the class conditional covariances. This more general model is known as 'Quadratic Discriminant Analysis' (QDA). Finally you are asked to consider how these models differ from the Naive Bayes model which we discussed in the lectures. Note throughout how different model assumptions imply different discriminant boundaries and hence different classifiers.*

   (a) ***[2 marks]***
   Describe the generative approach to classification. How does it differ from the discriminative approach?

   (b) ***[3 marks]***
   Derive the Bayes Optimal Classifier for binary classification, assuming misclassification loss.

   (c) ***[10 marks]***
   In a binary classification setting, assume that classes are distributed according to a Bernoulli random variable, $\mathcal{Y}$, whose outcomes are y, i.e. $y \sim \text{Bern}(\theta)$, where $\theta = p_{\mathcal{Y}}(y = 1)$. Furthermore we model the class conditional probability distributions for the random variable, $\boldsymbol{\mathcal{X}}$, whose outcomes are given by instances of particular input attribute vectors, $\mathbf{x} = [x_1, x_2, ..., x_m]^T \in \mathbb{R}^m$, as (note that here we will take care of the bias parameter explicitly, hence the absence of a leading '1' in the attribute vector):

$$\mathbf{x}|(y=0) \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \quad \text{where: } \boldsymbol{\mu}_0 \in \mathbb{R}^m,\ \boldsymbol{\Sigma}_0 \in \mathbb{R}^{m\times m}, \boldsymbol{\Sigma}_0^T = \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_0 \succ 0$$
$$\mathbf{x}|(y=1) \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \quad \text{where: } \boldsymbol{\mu}_1 \in \mathbb{R}^m,\ \boldsymbol{\Sigma}_1 \in \mathbb{R}^{m\times m}, \boldsymbol{\Sigma}_1^T = \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_1 \succ 0$$

   The off-diagonal elements of $\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1$ are not necessarily zero.
   Assume that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$ and show that the discriminant boundaries between the classes can be described by the following expression (you should clearly express $\mathbf{w}$ and b):

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \quad \text{where: } \mathbf{w} \in \mathbb{R}^m \text{ and } b \in \mathbb{R}$$

   (d) ***[2 marks]***
   What does this expression describe? Explain.

   (e) ***[4 marks]***
   Now assume that $\boldsymbol{\Sigma}_0 \neq \boldsymbol{\Sigma}_1$. What happens to the discriminant boundaries? Explain.

   (f) ***[4 marks]***
   Explain how this approach differs from that of Naïve Bayes.

2. *This problem focuses on discriminative classification. You begin by considering the Logistic Noise Latent Variable model and use it to motivate the Logistic Regression model, as we do in lectures. Following this you are asked to consider whether changing the parameterisation of the underlying logistic noise will imply a different classification model (it won't!). Next you are asked to repeat this analysis but for a Gaussian Latent Variable model. The resulting classification model is know as 'probit regression'. While it is similar in form to logistic regression, the probit function is less easy to manipulate than the logistic sigmoid, and furthermore has more sensitivity to outliers. Finally you are asked to consider a multinomial extension of the logistic regression model, and in particular to examine the form of the boundaries which exist between classes for this model.*

(a) ***[2 marks]***
Describe the discriminative approach to classification. How does it differ from the generative approach?

(b) ***[3 marks]***
Recall that in binary logistic regression, we seek to learn a mapping characterised by the weight vector, $\mathbf{w}$ and drawn from the function class, $\mathcal{F}$:

$$\mathcal{F} = \left\{ f_{\mathbf{w}}(\mathbf{x}) = \mathbb{I}\left[p_{\mathcal{Y}}(y=1|\mathbf{x}) > 0.5\right] \middle| p_{\mathcal{Y}}(y=1|\mathbf{x}) = \frac{1}{1+e^{-\mathbf{w}^\top \mathbf{x}}}, \mathbf{w} \in \mathbb{R}^{m+1} \right\}$$

Here $p_{\mathcal{Y}}(y|\mathbf{x})$ is the posterior output class probability associated with a data generating distribution, $\mathcal{D}$, which is characterised by the joint distribution $p_{\boldsymbol{\mathcal{X}},\mathcal{Y}}(\mathbf{x}, y)$. Provide a motivation for this form of the posterior output class probability $p_{\mathcal{Y}}(y|\mathbf{x})$ by considering a Logistic Noise Latent Variable Model. Remember that the noise in such a model characterises a random variable ϵ, with outcomes, ε, which are drawn i.i.d. as follows:

$\varepsilon \sim \text{Logistic}(a, b)$ where $w \in \mathbb{R}, b \in$

The characteristic probability distribution function for such a variable is:

$$p_\epsilon(\varepsilon|a,b) = \frac{\exp\left(-\frac{(\varepsilon-a)}{b}\right)}{b\left(1+\exp\left(-\frac{(\varepsilon-a)}{b}\right)\right)^2}$$

(c) ***[3 marks]***
If we allow the Logistic parameters to take general values $a \in \mathbb{R}$, $b > 0$ explain the effect which this has on the final logistic regression model.

(d) ***[4 marks]***
Let us assume instead a Gaussian Noise Latent Variable Model. Now ε is drawn i.i.d. as follows:

$$\varepsilon \sim \mathcal{N}(0,1)$$

Derive an expression for the posterior output class probability $p_{\mathcal{Y}}(y|\mathbf{x})$ in this case.

(e) ***[3 marks]***
How will the treatment of outliers in the data differ for these two models? Explain.

(f) ***[2 marks]***
For K-class multinomial regression, assuming misclassification loss, we can express a discriminative model for the posterior output class probability as:

$$p_{\mathcal{Y}}(y = j|\mathbf{x}) = \frac{\exp(\mathbf{w}_j \cdot \mathbf{x})}{\sum_{k=1}^{K} \exp(\mathbf{w}_k \cdot \mathbf{x})}$$

Where now $y \in \{1, ..., K\}$
Demonstrate that this model reduces to logistic regression when $K = 2$.

(g) ***[3 marks]***
For $K > 2$ derive an expression for the discriminant boundaries between classes. What does this expression describe?