Learning from Data, Tutorial Sheet for week 6

School of Informatics, University of Edinburgh

Instructor: Amos Storkey

1. A Naive Bayes Classifier for binary attributes $x_i \in \{0, 1\}$ is parameterised by $\theta_i^1 = p(x_i = 1|class = 1)$, $\theta_i^0 = p(x_i = 1|class = 0)$, and $p_1 = p(class = 1)$ and $p_0 = p(class = 0)$. Show that the decision boundary to classify a datapoint $\boldsymbol{x}$ can be written as $\boldsymbol{w}^T\boldsymbol{x} + b > 0$, and state explicitly $\boldsymbol{w}$ and b as a function of $\boldsymbol{\theta}^1, \boldsymbol{\theta}^0, p_1, p_0$.

2. Given a dataset $\{(\boldsymbol{x}^\mu, c^\mu), \mu = 1, \dots, P\}$, where $c^\mu \in \{0, 1\}$, logistic regression uses the model $p(c = 1|\boldsymbol{x}) = \sigma(\boldsymbol{w}^T\boldsymbol{x} + b)$. Assuming that the data is drawn independently and identically, show that the derivative of the log likelihood L of the data is

$$\nabla_{\boldsymbol{w}} L = \sum_{\mu=1}^{P} \left(c^\mu - \sigma\left(\boldsymbol{w}^T\boldsymbol{x}^\mu + b\right)\right)\boldsymbol{x}^\mu$$

3. Consider a dataset $\{(\boldsymbol{x}^\mu, c^\mu), \mu = 1, \dots, P\}$, where $c^\mu \in \{0, 1\}$, and $\boldsymbol{x}$ is a N dimensional vector.

- Show that if the training data is linearly separable with the hyperplane $\boldsymbol{w}^T\boldsymbol{x} + b$, the data is also separable with the hyperplane $\tilde{\boldsymbol{w}}^T\boldsymbol{x} + \tilde{b}$, where $\tilde{\boldsymbol{w}} = \lambda\boldsymbol{w}$, $\tilde{b} = \lambda b$ for any scalar $\lambda > 0$.
- What consequence does the above result have for maximum likelihood training of linearly separable data?

4. Consider a dataset $\{(\boldsymbol{x}^\mu, c^\mu), \mu = 1, \dots, P\}$, where $c^\mu \in \{0, 1\}$, and $\boldsymbol{x}$ is a P dimensional vector. (Hence we have P datapoints in a P dimensional space). If we are to find a hyperplane (parameterised by $(\boldsymbol{w}, b)$) that linearly separates this data we need, for each datapoint $\boldsymbol{x}^\mu$,

$$\boldsymbol{w}^T\boldsymbol{x}^\mu + b = \epsilon^\mu$$

where $\epsilon^\mu > 0$ for $c^\mu = 1$ and $\epsilon^\mu < 0$ for $c^\mu = 0$.

- Show that, provided that the data $\{\boldsymbol{x}^\mu, \mu = 1, \dots, P\}$ are linearly independent, a solution $(\boldsymbol{w}, b)$ always exists for any chosen values ϵ^μ.
- Discuss what bearing this has on the fact that the 600 handwritten digit training points are linearly separable in a 784 dimensional space.
- (Difficult) Comment on the relation between maximum likelihood training and the algorithm suggested above.