Machine Learning and Pattern Recognition, Tutorial Sheet Number 3

School of Informatics, University of Edinburgh, Instructor: Chris Williams

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

$$\nabla_{\boldsymbol{w}} L = \sum_{n=1}^{N} \left(y^n - \sigma\left(\boldsymbol{w}^T\boldsymbol{x}^n + b\right)\right)\boldsymbol{x}^n.$$

HINT: show that

$$\frac{d\sigma(z)}{dz} = \sigma(z)(1 - \sigma(z)).$$

2. Consider a dataset $\{(\boldsymbol{x}^n, y^n), n = 1, \ldots, N\}$, where $y^n \in \{0, 1\}$, and $\boldsymbol{x}$ is a D dimensional vector.

(a) Data is linearly separable if the two classes can be completely separated by a hyperplane. 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 $\lambda > 0$.

(b) What consequence does the above result have for maximum likelihood training of logistic regression for linearly separable data?

3. Consider a Bayesian linear regression model. Let

$$\begin{aligned} y &= mx + \eta \\ \eta &\sim \mathcal{N}(0, \sigma^2) \\ m &\sim \mathcal{N}(0, \tau^2) \end{aligned}$$

Assume that σ^2 and τ^2 are known. Note that to simplify the problem we have assumed that there is no x intercept. Identify the distributions of the following quantities under this model. (Merely identifying the family of distribution and its parameters is fine, e.g. Uniform$(0, \tau)$. You do not need to write down the pdf.)

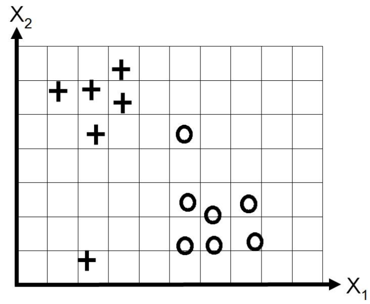
(a) What is $p(y|x = 1)$?

(b) Let y_1 equal the value of y when $x = 1$, i.e., $y_1 = m + \eta$. What is the joint distribution $p(y_1, m)$? Hint: Use the following facts

- For any random variable Z, we have $\mathrm{Var}(Z) = E[Z^2]$ when $E[Z] = 0$.
- For any random variables Y and Z, if Y and Z are independent, $\mathrm{Cov}(Y, Z) = 0$.
- For any random variables Y and Z, if $E[Y] = 0$ and $E[Z] = 0$, then $\mathrm{Cov}(Y, Z) = E[YZ]$.

(c) What is the posterior $p(m|y_1 = 1)$? Hint: Use what we did in Tutorial 1 with the bivariate Gaussian.

4. (Murphy, 8.7) Consider the following data set



(a) Suppose that we fit a logistic regression model, i.e., $p(y = 1|\boldsymbol{x}, \boldsymbol{w}) = \sigma(w_0 + w_1 x_1 + w_2 x_2)$. Suppose we fit the model by maximum likelihood, i.e., we minimize

$$J(\boldsymbol{w}) = -\ell(\boldsymbol{w}, \mathcal{D}_{\text{train}}),$$

where $-\ell$ is the logarithm of the likelihood above. Suppose we obtain the parameters $\hat{\boldsymbol{w}}$. Sketch a possible decision boundary corresponding to $\hat{\boldsymbol{w}}$.

Is your answer unique? How many classification errors does your method make on the training set?

(b) Now suppose that we regularize only the w_0 parameter, i.e., we minimize

$$J_0(\boldsymbol{w}) = -\ell(\boldsymbol{w}, \mathcal{D}_{\text{train}}) + \lambda w_0^2.$$

Suppose λ is a very large number, so we regularize w_0 all the way to 0, but all other parameters are unregularized. Sketch a possible decision boundary. How many classification errors does your method make on the training set? Hint: consider the behaviour of simple linear regression, $w_0 + w_1 x_1 + w_2 x_2$ when $x_1 = x_2 = 0$.

(c) Now suppose that we regularize only the w_1 parameter, i.e., we minimize

$$J_1(\boldsymbol{w}) = -\ell(\boldsymbol{w}, \mathcal{D}_{\text{train}}) + \lambda w_1^2.$$

Again suppose λ is a very large number. Sketch a possible decision boundary. How many classification errors does your method make on the training set?

(d) Now suppose that we regularize only the w_2 parameter, i.e., we minimize

$$J_2(\boldsymbol{w}) = -\ell(\boldsymbol{w}, \mathcal{D}_{\text{train}}) + \lambda w_2^2.$$

Again suppose λ is a very large number. Sketch a possible decision boundary. How many classification errors does your method make on the training set?