MLPM Tutorial 5

November 12, 2019

1. Maximum likelihood logistic regression maximises the log probability of the labels

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i, \mathbf{w})$$

with respect to the weight vector $\mathbf{w}$. The training data is said to be *linearly separable* if the two classes can be completely separated by a hyperplane. This means we can find a decision boundary

$$p(y_i = 1|\mathbf{x}_i, \mathbf{w}, w_0) = \sigma(\mathbf{w}^T\mathbf{x}_i + w_0) = 0.5$$

where σ is the sigmoid function, such that all points with $y = 1$ are on one side (i.e. have probability greater that 0.5) and all points $y \neq 1$ are on the other side of the hyperplane (i.e. have probability smaller than 0.5).

   (a) Show that if the training data is separable by a hyperplane specified by $\mathbf{w}$ and w_0 then it is also by the boundary $\lambda(\mathbf{w}^T\mathbf{x} + w_0)$ for $\lambda > 0$.

   (b) Show that for logistic regression likelihood maximisation results in $||\mathbf{w}||_2 \to \infty$.

2. Consider the data set in Figure 1.

   (a) Suppose that we fit a logistic regression model

$$p(y = 1|\mathbf{x}, \mathbf{w}) = \sigma(w_0 + w_1x_1 + w_2x_2)$$

   by maximum likelihood, obtaining parameters $\hat{\mathbf{w}}$. Sketch a possible decision boundary corresponding to $\hat{\mathbf{w}}$. Is your answer unique? How many classification errors does your solution make?

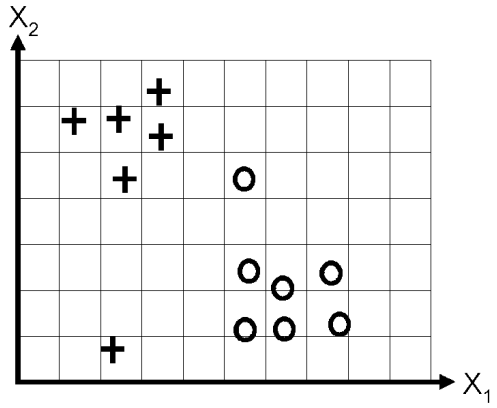

Figure 1: Example dataset.

(b) Now suppose that we regularise only the w_0 parameter, that is, we minmise

$$J_0(\mathbf{w}) = -\ell(\mathbf{w}) + \lambda w_0^2$$

where ℓ is the log-likelihood of $\mathbf{w}$. Sketch the decision boundary if $\lambda \to \infty$.

(c) Now suppose that we regularise only the w_1 parameter, that is, we minmise

$$J_1(\mathbf{w}) = -\ell(\mathbf{w}) + \lambda w_1^2$$

Sketch the decision boundary if $\lambda \to \infty$.

(d) Now suppose that we regularise only the w_2 parameter, that is, we minmise

$$J_2(\mathbf{w}) = -\ell(\mathbf{w}) + \lambda w_2^2$$

Sketch the decision boundary if $\lambda \to \infty$.

3. Let $E(\mathbf{w})$ be a differentiable function. Consider the gradient descent procedure

$$\mathbf{w}^{i+1} \leftarrow \mathbf{w}^i - \eta \nabla_{\mathbf{w}} E$$

where $\mathbf{w}^i$ is the estimate of $\mathbf{w}$ at the i^{th} iteration and η is the learning rate.

(a) Is it always the case that $E(\mathbf{w}^{i+1}) \leq E(\mathbf{w}^i)$?

(b) There always exists a value of η for which $E(\mathbf{w}^{i+1}) < E(\mathbf{w}^i)$?

4. A photon counter is pointed at a remote star for one minute, in order to infer the rate of photons arriving at the counter per minute, λ. Assuming the number of photons collected r has a Poisson distribution given by

$$p(r|\lambda) = e^{-\lambda}\frac{\lambda^r}{r!}$$

and assuming the prior $p(\lambda) \propto \frac{1}{\lambda}$, make Laplace approximations to the posterior distribution over λ.

Reparameterise the model such that $\ell = \log \lambda$ and find the Laplace approximation to the posterior over ℓ. Remember that the probability mass should be preserved when applying the transformation (see section 2.6.2 from Murphy).

5. Load the height/weight data from `data/heightWeightData.csv`. The first colum is the class label (1 = male, 2 = female), the second column is height (inches), the third weight (lbs). Implement a logistic regression classifier with $\ell 2$-regularisation in your favourite programming language. Use steepest gradient descent to find $\hat{\mathbf{w}}_{MAP}$ and plot the data and the predictive distribution $p(y = 1|\mathbf{x}, \hat{\mathbf{w}}_{MAP})$. Use 70% of the data for training and 30% for testing. Evaluate the classifier performance by calculating the misclassification, precision and recall rates as well as the F1 score.