Machine Learning and Pattern Recognition, Tutorial Sheet Number 3 Answers

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

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

Solution: $\mathcal{L}(\boldsymbol{w}, b) = \sum_{n=1}^{N} y^n \log \sigma\left(b + \boldsymbol{w}^T\boldsymbol{x}^n\right) + (1 - y^n)\log\left(1 - \sigma\left(b + \boldsymbol{w}^T\boldsymbol{x}^n\right)\right)$

Using $\nabla_{\boldsymbol{w}}\sigma(y) = (1 - \sigma(y))\sigma(y)\nabla_{\boldsymbol{w}} y$

$$\begin{aligned}
\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}, b) &= \sum_{n=1}^{N} y^n \frac{\nabla_{\boldsymbol{w}}\sigma(\cdot)}{\sigma(\cdot)} + \frac{\nabla_{\boldsymbol{w}}(1 - \sigma(\cdot))}{1 - \sigma(\cdot)} - y^n \frac{\nabla_{\boldsymbol{w}}(1 - \sigma(\cdot))}{1 - \sigma(\cdot)} \\
&= \sum_{n=1}^{N} y^n(1 - \sigma(\cdot))\boldsymbol{x}^n - \sigma(\cdot)\boldsymbol{x}^n + y^n\sigma(\cdot)\boldsymbol{x}^n \\
&= \sum_{n=1}^{N} y^n\boldsymbol{x}^n - y^n\sigma(\cdot)\boldsymbol{x}^n - \sigma(\cdot)\boldsymbol{x}^n + y^n\sigma(\cdot)\boldsymbol{x}^n \\
&= \sum_{n=1}^{N} \left(y^n - \sigma\left(\boldsymbol{w}^T\boldsymbol{x}^n + b\right)\right)\boldsymbol{x}^n
\end{aligned}$$

■

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?

Solution: The hyperplane $\tilde{b} + \tilde{\boldsymbol{w}}^T \boldsymbol{x} = \lambda\, b + \lambda\, \boldsymbol{w}^T \boldsymbol{x} \Rightarrow \lambda(b + \boldsymbol{w}^T \boldsymbol{x}) = 0$ is geometrically the same as $b + \boldsymbol{w}^T \boldsymbol{x} = 0$

If the data is linearly separable, the weights will continue to increase during the maximum likelihood training, and the classifications will become extreme (i.e. predictive probabilities of 0 or 1).

■

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

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

Solution:

(a) y is Gaussian because $y = mx + \eta$, and m, η are jointly Gaussian, and any linear combination of a Gaussian random variables is also Gaussian. So we'll just compute the mean and variance of y. For the mean

$$E[y|x = 1] = E[mx + \eta|x = 1] = E[mx|x = 1] + E[\eta|x = 1] = 0$$

For the variance

$$
\begin{aligned}
\text{Var}(y|x = 1) &= \text{Var}(mx + \eta|x = 1) \\
&= \text{Var}(mx|x = 1) + \text{Var}(\eta|x = 1) \\
&= \tau^2 + \sigma^2
\end{aligned}
$$

Therefore, $p(y|x = 1) = \mathcal{N}(y; 0, \tau^2 + \sigma^2)$. Note that this is a predictive distribution, i.e., we have integrated out m. By being clever we were able to avoid computing the integral by hand.

(b) As $p(m)$ and $p(y_1|m)$ are both Gaussian, so is the joint distribution $p(y_1, m)$. Its mean is $\mu = (0, 0)^T$. We already know $\text{Var}(m)$ and we computed $\text{Var}(y_1)$ in the previous part. This leaves $\text{Cov}(y_1, m)$. We compute this using a combination of the definition of covariance and minor trickery:

$$
\begin{aligned}
\text{Cov}(y_1, m) &= E\left[(y_1 - Ey_1)(m - Em)\right] \\
&= E[y_1 m] \\
&= E[m(m + \eta)] \\
&= E[m^2 + \eta m] \\
&= \tau^2,
\end{aligned}
$$

where in the last line we use $Em^2 = \text{Var}(m)$ and the fact that $E[\eta m] = \text{Cov}(\eta, m) = 0$.

This gives us that $p(y_1, m)$ is Gaussian with mean $(0,0)^T$ and variance

$$\Sigma = \begin{pmatrix} \sigma^2 + \tau^2 & \tau^2 \\ \tau^2 & \tau^2 \end{pmatrix}$$

(c) Now that we have $p(y_1, m)$ the results from Tutorial 1 tells us how to compute a conditional of a multivariate Gaussian.

> Let X_1 and X_2 be Gaussian $\mathcal{N}(x|\mu, \Sigma)$ with
>
> $$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$
>
> Then $p(x_1|x_2)$ is Gaussian with mean $\mu^c_{1|2}$ and variance $v^c_{1|2}$, where
>
> $$\mu^c_{1|2} = \frac{\sigma_{12} x_2}{\sigma_2^2}$$
> $$v^c_{1|2} = \frac{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}{\sigma_2^2} = \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2}$$

(You will not be expected to memorize this.)

In particular, that $p(m|y_1 = 1)$ is Gaussian with mean

$$\frac{\tau^2}{\sigma^2 + \tau^2}$$

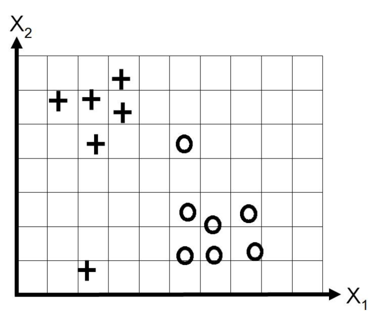
and variance

$$\tau^2 - \frac{\tau^4}{\sigma^2 + \tau^2}$$

As a sanity check, consider what happens as $\sigma^2 \to 0$. When that happens, our measurements get precise, so we should become certain about the slope m even from one data point. So the mean should converge to 1 and the variance to 0. Check this.

■

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?

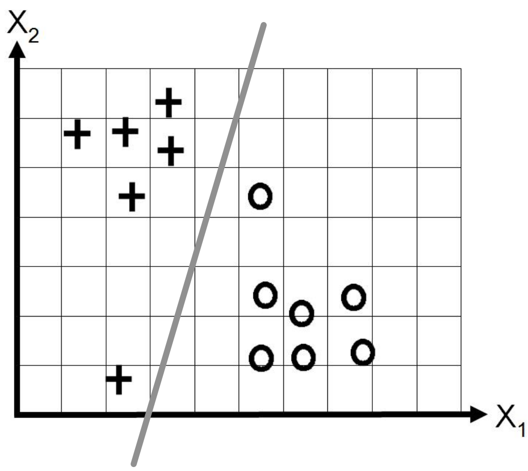
(d) Now suppose that we regularize only the w_2 parameter, i.e., we minimize

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

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?

Solution:

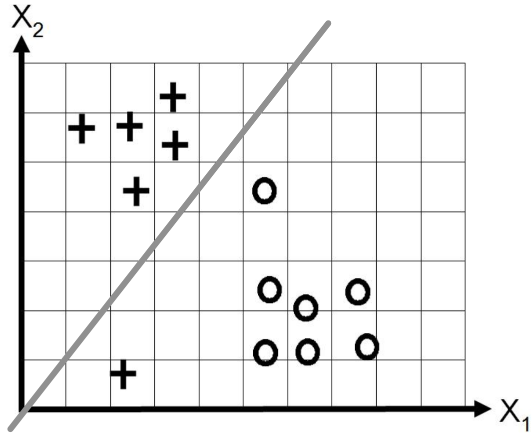
(a) As the data are linearly separable, logistic regression will find a line that fits the data perfectly. There will be no classification errors on the training set. The line is not unique (imagine wiggling



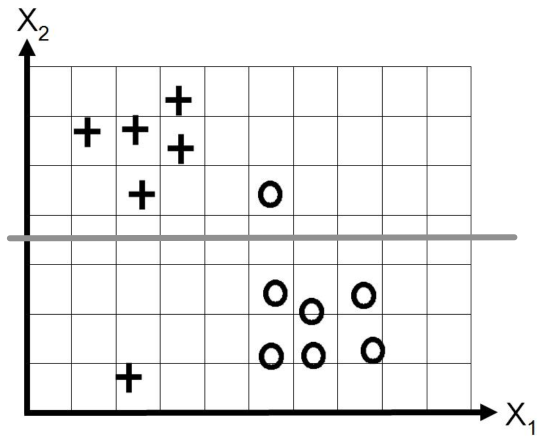
it).

(b) Since $w_0 = 0$, this means that the point $(0, 0)$ must be on the decision boundary, because at that point $\sigma(w_0 + w_1 x_1 + w_2 x_2) = \sigma(0) = 0.5$. So regularized logistic regression will find the best decision boundary that passes through $(0, 0)$. It will make one mistake on the training data.

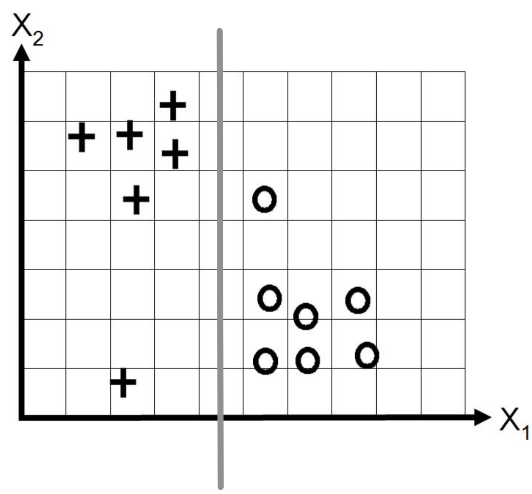
As an aside, it is for this reason that in regularized logistic or linear regression, we usually do *not* penalize the bias term (i.e., the weight that corresponds to the feature that is always 1).



(c) As the regularizer forces $w_1 = 0$, the decision boundary will be a horizontal line. There will be two classification errors.



(d) As the regularizer forces $w_2 = 0$, the decision boundary will be a vertical line. There will be zero classification errors.



■