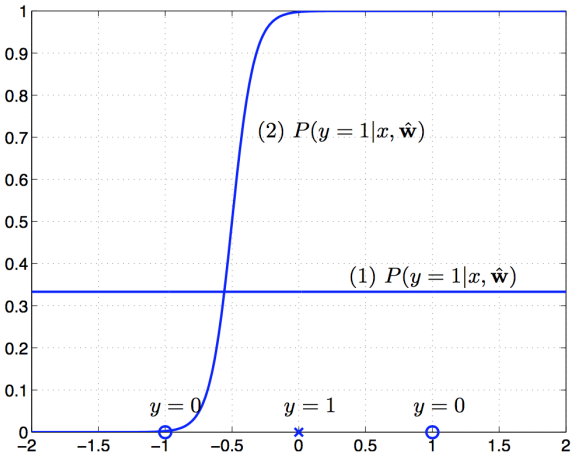
6.867: Exercises 3

Sept 29, 2017

Contents

1 Logistic regression: basic intuition with one-dimensional data



Consider a simple one dimensional logistic regression model

$$P(y = 1 \mid x, w) = \sigma(w_0 + w_1 x)$$

where $\sigma(z) = (1 + \exp(-z))^{-1}$ is the logistic function.

The figure above shows two possible conditional distributions $P(y = 1 \mid x, w)$, viewed as a function of x, that we can get by changing the parameters w.

Assume we have a data set $\mathcal{D} = \{(-1, 0), (0, 1), (1, 0)\}$.

1. Please indicate the number of classification errors for each conditional given $\mathcal{D}$.

   Conditional (1) makes ( ) classification errors

   Conditional (2) makes ( ) classification errors

2. Which of these two hypotheses assigns a higher likelihood to the data?

3. If your loss function for predictions was

$$L(g, a) = \begin{cases} 0 & \text{if } g = a \\ 1 & \text{if } g = 1 \text{ and } a = 0 \\ 10 & \text{if } g = 0 \text{ and } a = 1 \end{cases}$$

   - What output would you predict for $x = -1$ when conditional (1) is the result of your learning?
   - What output would you predict for $x = -1$ when conditional (2) is the result of your learning?

2 Logistic Regression: maximum likelihood

(Bishop 4.14) Show that for a linearly separable data set, the maximum likelihood solution for the logistic regression model is obtained by finding a vector w whose decision boundary $w^T \phi(x) = 0$ separates the classes and then taking the magnitude of w to infinity.

3 More on Logistic regression

We are interested in regularizing the terms separately in logistic regression.

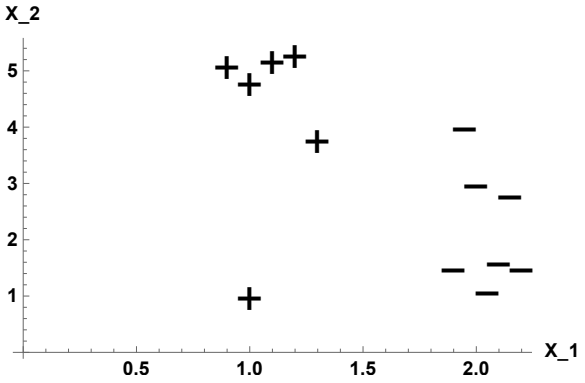
(a) Consider the data in the figure below where we fit the model

$$P(y = 1 \mid x, w) = \text{Sigmoid}(w_0 + w_1 x_1 + w_2 x_2)$$

Suppose we fit the model by maximum likelihood, that is, we minimize

$$J(w) = -\log \Pr(D_{\text{train}}; w)$$

Sketch a possible decision boundary corresponding to w^*.



(b) Is your decision boundary unique?

(c) How many classification errors does it make on the training set?

(d) Now suppose we regularize only the w_0 parameter; that is, we minimize

$$J(w) = -\log \Pr(D_{\text{train}}; w) + \lambda w_0^2$$

with λ approaching ∞. Sketch a possible decision boundary corresponding to w^*.

(e) How many classification errors does it make on the training set?

(f) Now suppose we regularize only the w_1 parameter; that is, we minimize

$$J(w) = -\log \Pr(D_{\text{train}}; w) + \lambda w_1^2$$

with λ approaching ∞. Sketch a possible decision boundary corresponding to w^*.

(g) How many classification errors does it make on the training set?

(h) Now suppose we regularize only the w_2 parameter; that is, we minimize

$$J(w) = -\log \Pr(D_{\text{train}}; w) + \lambda w_2^2$$

with λ approaching ∞.

(i) How many classification errors does it make on the training set?

4 SVMs: basic intuition with one-dimensional data

Assume that our training data is four 1-dimensional points, as follows:

index	x	y
1	-2	-1
2	-0.1	-1
3	0.1	1
4	1	1

(a) Find the values of all the α_i that would be found by the (linear) SVM training algorithm. You should be able to do this without going through the Lagrangian minimization procedure. Think about the conditions for the optimization directly.

(b) What would the offset be for these values of α_i?

(c) What if the value of C were set to 1? What would happen to the values of α_i and the offset? Explain.

5 SVMs: Where's the hyperplane?

(Bishop 7.3) Show that, irrespective of the dimensionality of the data space, a data set consisting of just two data points, one from each class, is sufficient to determine the location of the maximum-margin hyperplane.

6 SVMs: the α_i and $\|w\|$

Here we investigate the relation between the SVM dual variables and the margin. That is, we want to discover the relationship between $\sum_i \alpha_i$ and $1/\|w\|$.

Show that: $\sum_i \alpha_i = \text{margin}^{-2}$, where the α_i are the optimal solution to the primal problem. (Hint: Recall that margin = $1/\|w\|$)

7 Multi-class SVMs

The multi-class SVM generalizes the binary SVM to multi-class classification. This involves introducing a weight vector $\vec{w}^{(k)}$ and $b^{(k)}$ for each class $k = 1, \ldots, K$ (where K is the number of classes). Learning solves the following optimization problem, where there is still only one slack variable ξ_j for each data point, but now there are $K - 1$ constraints per data point:

$$\min_{\{\vec{w}^{(k)}, b^{(k)}\}} \sum_{k=1}^{K} \|\vec{w}^{(k)}\|_2^2 + C \sum_j \xi_j$$

subject to

$$\begin{aligned} \vec{w}^{(y_j)} \cdot \vec{x}_j + b^{(y_j)} &\geqslant \vec{w}^{(k)} \cdot \vec{x}_j + b^{(k)} + 1 - \xi_j \qquad \forall j \text{ and } k \neq y_j \\ \xi_j &\geqslant 0 \qquad \forall j. \end{aligned}$$

Prediction for a new data point $\vec{x}$ is performed using the rule

$$\hat{y} \leftarrow \arg\max_k \ \vec{w}^{(k)} \cdot \vec{x} + b^{(k)}.$$

This problem compares the binary prediction rule $\text{sign}(\vec{w} \cdot \vec{x} + b)$ to the multi-class prediction rule in the case that $K = 2$, and shows how to reduce between the two of them.

1. Demonstrate $\vec{w}$ and b as a function of $\vec{w}^{(1)}$, $b^{(1)}$, $\vec{w}^{(2)}$ and $b^{(2)}$ such that the predictions made for all data points $\vec{x}$ using the new binary prediction rule are the same as what would have been made using the multi-class prediction rule with $\vec{w}^{(1)}$, $b^{(1)}$, $\vec{w}^{(2)}$, $b^{(2)}$.

2. Next you should show the converse. Given $\vec{w}$ and b, demonstrate $\vec{w}^{(1)}$, $b^{(1)}$, $\vec{w}^{(2)}$ and $b^{(2)}$ (as a function of $\vec{w}$ and b) such that the predictions made for all data points $\vec{x}$ using the multi-class prediction rule are the same as what would have been made using the binary prediction rule with $\vec{w}$ and b.