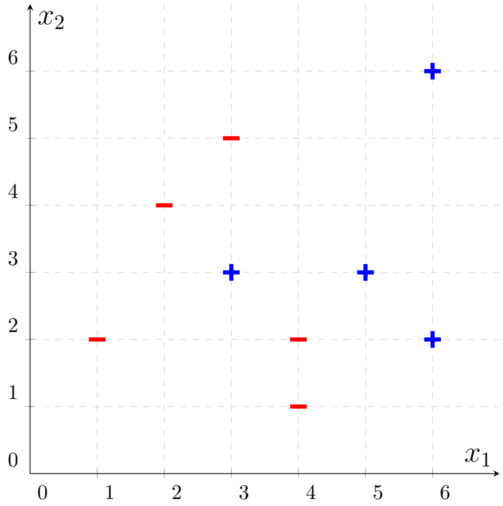
6.390 Introduction to Machine Learning

Recitation Week #4

Issued September 25, 2023

1. Consider the training data set plotted below:



Assume that $x^{(1)}, \ldots, x^{(n)}$ are the coordinate vectors for the $n = 9$ points shown on the figure, $y^{(1)}, \ldots, y^{(n)}$ are the corresponding labels (for example, $y^{(i)} = 1$ for the "+" points, $y^{(i)} = -1$ for the "-" points), and $h : \mathbb{R}^2 \to \{+1, -1\}$ is a hypothesis for mapping the data point coordinates to the binary labels.

(a) Find the classification accuracy of the hypothesis

$$h\left(\begin{bmatrix} x1 \\ x2 \end{bmatrix}\right) = \begin{cases} +1, & \text{if } x_1 + x_2 - 7 > 0, \\ -1, & \text{if } x_1 + x_2 - 7 \leq 0, \end{cases}$$

on this training set. (See Section 4.6 in the Lecture Notes.)

(b) Draw a hyperplane which defines a linear classifier that obtains the smallest training error in terms of highest classification accuracy. Be sure to also draw the normal vector. Specify values for θ, θ_0 that define the hyperplane you drew.

(c) Suppose we remove the datapoint $([3,3]^\top, +1)$. Let two hypotheses be considered different if there exists a test point (i.e., not necessarily from the data set shown) that they would classify differently. How many different hypotheses are there that obtain zero training error? Explain your answer.

(d) In the set of solutions with zero training error, is there a reason to prefer any specific hypotheses over others? How might we find one or more of those?

(e) Suppose that we learn the parameters of a linear logistic classifier $h(x) = \sigma(\theta^\top x + \theta_0)$ by minimizing the logistic regression objective function with regularization,

$$J(\theta, \theta_0) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{\text{nll}}(\sigma(\theta^\top x^{(i)} + \theta_0), y^{(i)}) + \lambda \|\theta\|^2.$$

How does this approach address the issue of differentiating between linear separators discussed earlier in this problem?

2. For each of the following, determine if the statement is true or false and justify your reasoning.

   (a) If we take any linearly separable data set and add a new feature, it is still guaranteed to be linearly separable.

   ○ True ○ False

   (b) If we take any linearly separable data set and remove a feature, it is still guaranteed to be linearly separable.

   ○ True ○ False

   (c) If we take any data set that is not linearly separable and remove a feature, it is still guaranteed to not be linearly separable.

   ○ True ○ False

   (d) If we take any data set that is not linearly separable and remove a data point, it is still guaranteed to not be linearly separable.

   ○ True ○ False

3. Beatriz used logistic regression on a data set derived from people living in Massachusetts to learn a linear logistic classifier $\sigma(\theta^\top x + \theta_0)$ giving the probability that an adult with features x will develop heart disease in the next decade. Let θ^* and θ_0^* be the set of parameters that Beatriz learns.

   Her friend, John, would like to use the same logistic regression classifier (i.e., with the θ^* and θ_0^* learned by Beatriz) to make predictions for people living in Norway. However, he notices that heart disease is much less common in Norway and thinks that the model may need to be adjusted to account for this. For now, he decides to follow Beatriz in using logistic regression and decides to consider only changes to the offset θ_0 (that is, he always uses the same $\theta = \theta^*$ as Beatriz).

   (a) Consider a specific patient with feature vector x. How could John adjust the offset parameter, relative to the θ_0^* learned by Beatriz, so as to make smaller the predicted probability of this patient developing heart disease?

   (b) John realizes that choosing the right value of θ_0 is tricky since he doesn't have access to any labeled data from Norway. John tells Beatriz that he only plans to use the model to find the 10% of individuals with highest probability of developing heart disease so that he can closely follow them and make sure they are tested appropriately. "Aha!", says Beatriz. "In that case, any value of θ_0 would suffice, and you can simply make use of my original linear logistic classifier!" Is Beatriz correct? Why or why not?