6.867: Exercises (Week 2)

Sept 22, 2017

1 Decreasing Variance (Bishop 3.11)

We have seen that, as the size of a data set increases, the uncertainty associated with the posterior distribution over model parameters decreases. The variance $\sigma_N^2(x)$ of the predictive distribution (see page 18 of lecture notes 4) associated with the linear regression function is given by

$$\sigma_N^2(x) = \sigma^2 + x^T S_N x.$$

Suppose we add a new data point x_{N+1} to the data set. Make use of the matrix identity (Sherman-Morrison formula)

$$(M + vv^T)^{-1} = M^{-1} - \frac{(M^{-1}v)(v^T M^{-1})}{1 + v^T M^{-1} v}$$

and

$$S_{N+1}^{-1} = S_N^{-1} + \sigma^{-2} x_{N+1} x_{N+1}^T$$

to show that

$$\sigma_{N+1}^2(x) \leqslant \sigma_N^2(x).$$

2 Bayesian Linear Regression with Advertisement Data

Recall Bayesian Linear Regression with a Gaussian Prior as a way to understand Ridge Regression. We shall understand the effect of this in the context of Advertisement Data.

Consider the setting where sales is the target variable and the features are spending on TV, Radio and Newspaper. Use a Gaussian prior on model parameters with mean 0 and covariance matrix as the identity.

We want to understand the effect of having more and more data points on the posterior. Calculate the trace of co-variance matrix for the model parameters as number of data points increases and observe that it is decreasing.

3 Convex Hull and Linearly Separability (Bishop 4.1)

Given a set of data points $\{x_n\}$, we can define the *convex hull* to be the set of all points x given by

$$x = \sum_n \alpha_n x_n$$

where $\alpha_n \geqslant 0$ and $\sum_n \alpha_n = 1$. Consider a second set of points $\{y_n\}$ together with their corresponding convex hull. By definition, the two sets of points will be linearly separable if there exists a vector $\hat{w}$ and a scalar w_0 such that $\hat{w}^T x_n + w_0 > 0$ for all x_n, and $\hat{w}^T y_n + w_0 < 0$ for all y_n. Show that if their convex hulls intersect, the two sets of points cannot be linearly separable, and conversely that if they are linearly separable, their convex hulls do not intersect.

4 Perceptron

⋆ Stanford CS246 Winter 2017

You are given information about five patients who are being tested for diabetes. The results of Fasting Plasma Glucose (FPG) test and the Oral Glucose Tolerance (OGT) test are provided for each of these patients. The outcome variable indicating if each patient has been diagnosed with diabetes (+1) or not (-1) is also available in the data.

We already generated features for each patient x as follows:

- $\phi_1(x) = 1$ if FPG test is positive, otherwise it is set to 0.
- $\phi_2(x) = 1$ if OGT test is positive, otherwise it is set to 0.
- $\phi_3(x) = -1$, a bias term.

Given a weight vector $w = (w_1, w_2, w_3)$, our classifier returns +1 if $w_1\phi_1(x) + w_2\phi_2(x) + w_3\phi_3(x) > 0$ and -1 otherwise. Our training set comprises of the following features and labels in Table 1:

PatientID	ϕ_1	ϕ_2	ϕ_3	Label
1	0	0	-1	+1
2	1	1	-1	+1
3	1	1	-1	+1
4	0	1	-1	-1
5	1	0	-1	-1

Table 1: **Diabetes Dataset**

1. Compute the first four updates of the Perceptron training algorithm using the diabetes data provided in Table 1. Fill in the following table, using the given initial Perceptron weights $w = (w_1, w_2, w_3) = (0, 0, 0)$, $\eta = 1/5$ (η is the step size).

w	w_1	w_2	w_3
After Observing PatientID = 1			
After Observing PatientID = 2			
After Observing PatientID = 3			
After Observing PatientID = 4			

2. Will the Perceptron algorithm return a solution for the diabetes dataset shown in Table 1? Why?

3. Linear classifiers such as Perceptrons are often insufficient to represent a dataset using a given set of features. However, it is often possible to find new features using nonlinear functions of our existing features which do allow linear classifiers to separate the data. Nonlinear features result in more expressive linear classifiers. For example, consider the following data set, where $+$s represent positive examples and $-$s represent negative examples.

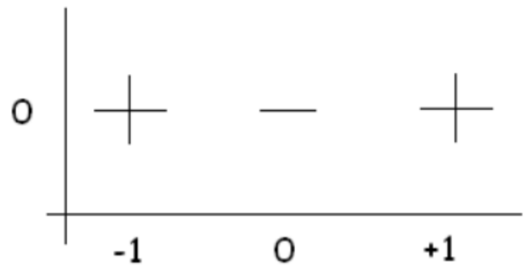

Figure 1: **The original data points are not linearly separable**

Let x_1 and x_2 denote the two dimensions of the data shown in Figure 1. No linear classifier can separate the positive examples $(-1, 0)$ and $(1, 0)$ from the negative example $(0, 0)$ in the dataset shown in Figure 1. Rather than using a single feature, if we perform a nonlinear mapping or transformation $\psi = (x_1^2, x_2 + 1)$, the positive examples are both mapped to $(1, 1)$ and the negative example is mapped to $(0, 1)$, and we see the data is now linearly separable (Figure 2).

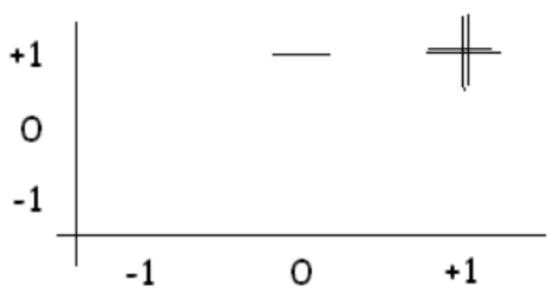

Figure 2: **After the non-linear transformation the data becomes linearly separable**

Which of the following transformations linearly separate the diabetes dataset shown in Table 1? Justify your answer about each transformation briefly. If a particular transformation linearly separates the data, also provide the value of w that separates the two classes in the data?

(i) $\psi = (\phi_1^2, \phi_1\phi_2, -1)$

(ii) $\psi = ((\phi_1 \text{ xor } \phi_2), \phi_2, -1)$, where a xor b is 1 if either $a = 1$ or $b = 1$ but not both.

(iii) $\psi = (\phi_1 - \phi_2, \phi_1 + \phi_2, -1)$.

5 Perceptron and Ordering

⋆ CMU 10-601 Spring 2015

Consider running the Perceptron algorithm on some sequence of examples S (an example is a data point and its label). Let S' be the same set of examples as S, but presented in a different order.

1. Does the Perceptron algorithm necessarily make the same number of mistakes on S as it does on S'?

2. If so, why? If not, show such an S and S' where the Perceptron algorithm makes a different number of mistakes on S' than it does on S.

6 The Worst Case of Perceptron

⋆ Shalev-Shwartz, Shai, and Shai Ben-David. Understanding machine learning: From theory to algorithms.

For any positive integer m, find a sequence of examples $\{(x_1, y_1), ..., (x_m, y_m)\}$ such that when running the perceptron on this sequence of examples starting from $w^{(0)} = 0$, it makes at least m updates before converging.
Hint: Try to use training data drawn from $\mathbb{R}^m$.