Homework 6, Due May 16

MPCS 53111 Machine Learning, Winter 2016
University of Chicago

For all questions in which you obtain an empirical result, such as a plot, briefly discuss the result: Is the result as expected? Does the result reveal some interesting aspect of the data or the algorithm.

Practice problems, do not submit

1. When the data for a SVM is not linearly separable, or we are willing to tolerate a few points inside the margin, a *soft margin* classifier is used. In this the optimization function has a *regularization term* as shown below:
$$\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{m} \max\left\{0, 1 - y^{(i)}(\mathbf{w}\cdot\mathbf{x} + b)\right\}$$
   (a) Describe in a few words how this change makes sense. (Hint: See §12.3.3, "Mining of Massive Datasets.")

   (b) `sklearn` uses an apparently different formulation. Explain why the two formulations are equivalent.

2. Read §9.5 in ISLR for a discussion on the relationship between Logistic Regression and SVMs and summarize it.

Graded problems, submit

3. The accompanying starter code implements an `sklearn`-style regressor class `KNNRegressor` for k-nearest-neighbor linear regression. It also generates data using the following underlying model—

   - $x \sim \text{Uniform}(0, 100)$.
   - $y(x) = 3 + 4x - 0.05x^2 + \epsilon$, in which $\epsilon \sim \text{Uniform}(-25, 25)$.

   Finally it plots the data points, the underlying function without the noise ($y(x) - \epsilon$), and the hypothesis function for k-nearest-neighbor linear regression.

(a) Similarly, write the class `LWRegressor` to implement locally weighted regression in which the kernel is the rbf_kernel that takes a user specified γ parameter.

(b) Create a combined plot that includes (i) data points, (ii) the underlying function without the noise, (iii) the hypothesis function for k-nearest-neighbor linear regression with $k = 5$, and (iv) the hypothesis function for locally weighted regression with $\gamma = 1/40$.

(c) Using crossvalidation on the above dataset, determine the best value of (i) k for k-nearest-neighbor linear regression, and (ii) γ for locally weighted regression, and clearly state those values. Which model do you recommend using for this artificial dataset?

4. I have a classification problem for SVMs with inputs x_1, x_2. I believe the data is linearly separable in the space of the following features: $x_i^3, x_i^2 x_j, x_i^2, x_i x_j, x_i$, in which i and j are distinct indices from $\{1, 2\}$. What feature vector should I use, if I want to use the kernel trick? What is the corresponding kernel function?

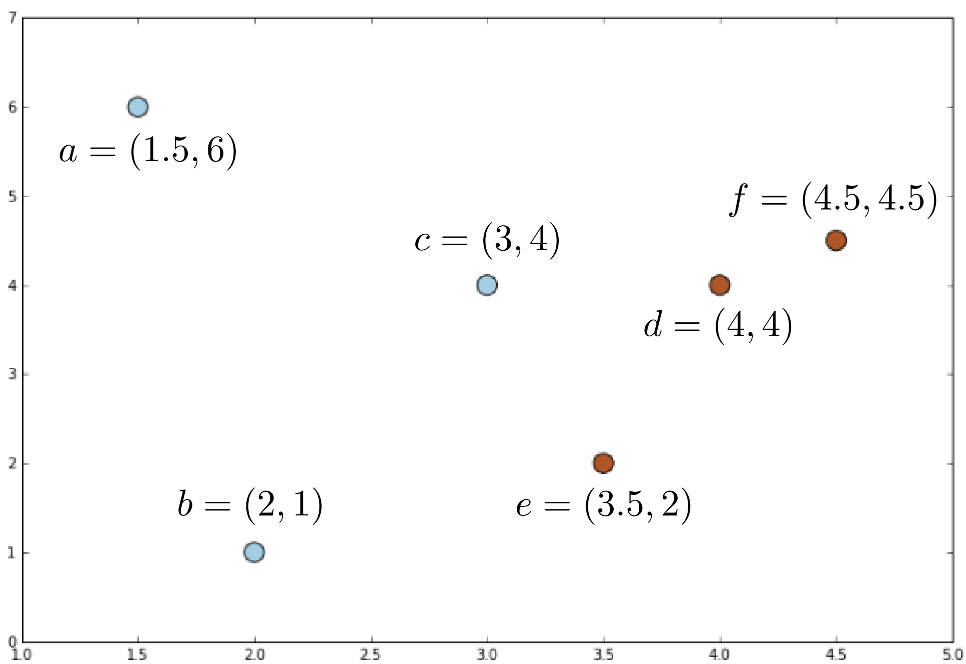

Figure 1: SVM classification problem.

5. In this problem you will get a feel for maximum margin separators. Refer to the dataset shown in Figure 1

   (a) Make your best guess on the optimal separator and determine the distance of the closest example to it. (There is no easy way to manually determine the maximum margin separator, so correctness is not required.)

   (b) Use `sklearn.svm.SVC` to find the maximum margin separator, and plot the corresponding separator and the points. (Use a linear kernel, and $C \geq 100$ to prevent a soft margin.) What is the distance of the closest example to the separator?

   (c) Repeat 5a and 5b but with point e at $(5, 2)$.

   (d) Repeat 5b but switch the class of f. Plot the soft margin separators (see 1) for $C \in \{0.1, 1, 10\}$.

   (e) Repeat 5d but with the kernels *polynomial* and *rbf*. These kernels use special parameters; you may have to change them to correctly classify all points. (For an intuitive explanation on why the rbf kernel works, see pg 16, Andrew Ng's notes on SVMs.)