6.867 Machine learning

Mid-term exam

October 14, 2009

(2 points) Your name and MIT ID:

Problem 1

Consider a simple classification problem (of the kind that you could only encounter in an exam). The training data consist of only three labeled points

$$(x_1 = -1, y_1 = 1), (x_2 = 0, y_2 = -1), (x_3 = +1, y_3 = +1)$$

which we will try to separate with a linear classifier through origin in the feature space. In other words, our discriminant function is of the form $\underline{\theta} \cdot \underline{\phi}(x)$. The corresponding primal and dual estimation problems are given by

$$\textbf{Primal:} \quad \text{Minimize} \quad \frac{1}{2}\|\underline{\theta}\|^2 + C\sum_{i=1}^{3}\xi_i$$
$$\text{subject to} \quad y_i(\underline{\theta} \cdot \underline{\phi}(x_i)) \geq 1 - \xi_i, \;\; \xi_i \geq 0, \;\; i = 1, 2, 3$$

$$\textbf{Dual:} \quad \text{Maximize} \quad \sum_{i=1}^{3}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{3}\alpha_i\alpha_j y_i y_j K(x_i, x_j)$$
$$\text{subject to} \quad 0 \leq \alpha_i \leq C, \;\; i = 1, 2, 3$$

1.1 (3 points) We decided to solve the problem in the dual using kernel $K(x, x') = 1 + (xx')^2$. What is the feature representation $\underline{\phi}(x)$ corresponding to this kernel?

1.2 (3 points) Are the training points linearly separable using this kernel?

1.3 (4 points) Suppose we selected any kernel $K_1(x, x')$ that makes the training points linearly separable. If the same points are not linearly separable using another kernel $K_2(x, x')$, what happens with the kernel $K(x, x') = K_1(x, x') + K_2(x, x')$? Are the points separable? Briefly justify your answer.

1.4 (2 points) If we decrease the slack penalty C, the solution might not satisfy the margin constraint for $(x_2 = 0, y_2 = -1)$ without a positive slack $\xi_2 > 0$. What does this mean in terms of α_2?

1.5 (4 points) Suppose $K(x, x') = 1 + (xx')^2 + (xx')^3$ and we select $C < 1$. Express the value of the discriminant function in the dual form for $x_2 = 0$. Will we necessarily have a positive slack $\xi_2 > 0$?

Problem 2

The Perceptron algorithm is perhaps the simplest way to solve classification problems. We also like the radial basis kernel. The kernel Perceptron algorithm with the radial basis kernel is given by

Algorithm 1

Initialize: $\alpha_1 = \ldots = \alpha_n = 0$

Cycle through $i = 1, \ldots, n$ until no mistakes

if $y_i(\sum_{j=1}^{n} \alpha_j y_j K(\underline{x}_j, \underline{x}_i)) \leq 0$, then $\alpha_i \leftarrow \alpha_i + 1$

where $K(\underline{x}, \underline{x}') = \exp(-\frac{1}{2\sigma^2}\|\underline{x} - \underline{x}'\|^2)$, $\sigma^2 > 0$.

2.1 (4 points) What can you say about the convergence of Algorithm 1? If convergence requires additional conditions, please provide them.

2.2 (6 points) Check all that apply

( ) a) The geometric margin of the solution that the algorithm finds (if it converges) depends on the order in which we cycle through the training examples.

( ) b) Suppose all the training inputs $\{\underline{x}_i\}_{i=1,\ldots,n}$ are distinct. Then, for a small enough kernel width σ, the algorithm must converge after at most n mistakes.

( ) c) The algorithm may converge before n mistakes for a larger value of σ^2

Briefly explain why you did/did not check b).

Problem 3

A friend of ours claimed that she can solve and use maximum margin linear classifiers with only access to code developed for anomaly detection. The training and testing routines she claims are sufficient are given below.

$$(\hat{\underline{\theta}}, \hat{\rho}) = \texttt{train}(\underline{\phi}_1, \ldots, \underline{\phi}_n): \quad \text{Minimize } \frac{1}{2}\|\underline{\theta}\|^2 - \rho \text{ subject to } \underline{\theta} \cdot \underline{\phi}_i \geq \rho, \;\; i = 1, \ldots, n \text{ Return } \hat{\underline{\theta}},\, \hat{\rho}$$

$$y = \texttt{test}(\underline{\theta}, \rho, \underline{\phi}): \quad \text{Return } +1 \text{ if } \underline{\theta} \cdot \underline{\phi} \geq \rho \text{ else return } -1$$

We were a bit uncertain about her claim but decided to try anyway. Let's start by finding the maximum margin linear classifier of the form

$$\hat{y} = \text{sign}\big(\, \underline{\theta} \cdot \underline{\phi}(\underline{x}) \,\big)$$

based on n training examples $\underline{\phi}(\underline{x}_1), \ldots, \underline{\phi}(\underline{x}_n)$ and ± 1 labels $y_1, \ldots, y_n$. Assume that the training set is linearly separable.

3.1 (4 points) What are the feature vectors that we should pass onto the training routine `train`?

3.2 (4 points) Let $\hat{\underline{\theta}}$ and $\hat{\rho}$ be the parameters returned by `train` with your choice of training feature vectors. What are the three arguments that we should give to `test` such that it would classify the test point $\underline{\phi}(\underline{x})$ in the same way as the maximum margin classifier?

3.3 (4 points) What is the geometric margin that the maximum margin classifier achieves on the training set? Express your answer in terms of $\hat{\underline{\theta}}$ and $\hat{\rho}$.

3.4 (4 points) Encouraged by this we thought that perhaps it is also possible to train a maximum margin linear classifier with an offset parameter, i.e., find a classifier of the form $\hat{y} = \text{sign}\big(\underline{\theta} \cdot \underline{\phi}(\underline{x}) + \theta_0\big)$ using just the two routines. Is this possible? Please justify your answer briefly.

Problem 4

We asked a few students to rate their midterm exam according to whether they thought it was difficult ($y = 1$), all right ($y = 2$), or easy ($y = 3$). Each student also provided us with a few pieces of information about themselves such as other courses they had taken, the program they were in, and so on. We could use this additional information to construct a feature vector $\underline{\phi}_i$ for each student $i = 1, \ldots, n$. On the basis of the rating labels, $y_1, \ldots, y_n$ and the feature vectors, $\underline{\phi}_1, \ldots, \underline{\phi}_n$, we could learn to predict how a particular type of student would react to the exam.

We decided to divide the prediction task into two binary classification tasks

Task 1: whether $y = 1$ (binary label -1) or $y > 1$ (binary label +1)

Task 2: whether $y \leq 2$ (binary label -1) or $y = 3$ (binary label +1)

So, we needed two binary classifiers. Since the ratings fall on an ordinal scale, it seemed wise to couple these tasks together. We opted to use common parameters $\underline{\theta}$ for the two tasks but different thresholds b_1 and b_2 for the task 1 and 2, respectively. The corresponding estimation problem is given by

$$\begin{aligned}
&\text{Minimize } \frac{1}{2}\|\underline{\theta}\|^2 \quad \text{with respect to } \underline{\theta},\ b_1, \text{ and } b_2, \text{ subject to} \\
&\quad \text{Task 1:} \quad -1(\underline{\theta}\cdot\underline{\phi}_i - b_1) \geq 1 \ \text{ if } y_i \leq 1, \quad +1(\underline{\theta}\cdot\underline{\phi}_i - b_1) \geq 1 \ \text{ if } y_i > 1 \\
&\quad \text{Task 2:} \quad -1(\underline{\theta}\cdot\underline{\phi}_i - b_2) \geq 1 \ \text{ if } y_i \leq 2, \quad +1(\underline{\theta}\cdot\underline{\phi}_i - b_2) \geq 1 \ \text{ if } y_i > 2 \\
&\text{for all } i = 1, \ldots, n
\end{aligned}$$

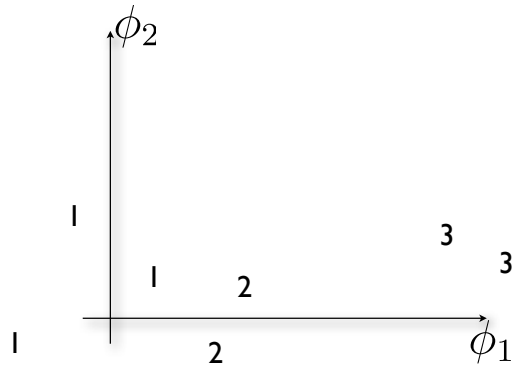

Figure 4.1. Student exam ratings in the feature space.

4.1 (3 points) If the problem is separable in the sense that the quadratic program has a solution, are we guaranteed that the solution $\hat{\underline{\theta}}$, $\hat{b}_1$, $\hat{b}_2$ satisfies $\hat{b}_1 \leq \hat{b}_2$?

4.2 (4 points) Suppose we omit Task 2 constraints altogether and only focus on Task 1 in order to solve for $\hat{\underline{\theta}}$ and $\hat{b}_1$. Draw the resulting decision boundary and margin constraints in Figure 4.1 based on the data in the figure.

4.3 (6 points) Under which conditions is the solution $\hat{\underline{\theta}}$ based on Task 1 alone also the solution to the combined task?

Additional set of figures

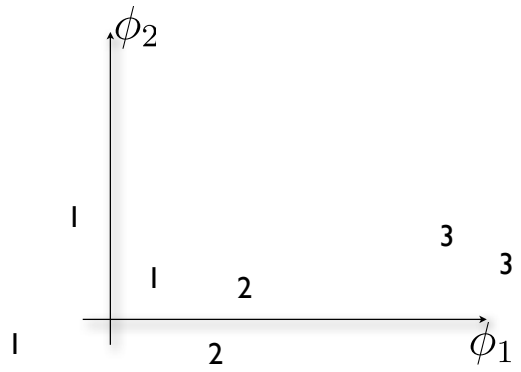

Figure 4.1. Student exam ratings in the feature space.