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

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

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

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

$\underline{\phi}(\underline{x}) = (1, |x|)^T$

1.2 3Using this kernel (feature mapping), are the three training examples linearly separable through origin in the feature space?

Y

1.3 (4 points) Consider any pair of kernels $K_1(x, x')$ and $K_2(x, x')$ such that the training points are linearly separable with $K_1(x, x')$ but not with $K_2(x, x')$. What happens when we add the two and use $K(x, x') = K_1(x, x') + K_2(x, x')$? Are the points separable now? Briefly justify your answer.

The points are separable. By adding kernels we concatenate the corresponding feature vectors. We can always set the parameters associated with the additional coordinates (here corresponding to K_2) to zero.

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?

$\alpha_2 = C$

1.5 (4 points) Assume $K(x, x') = 1 + |xx'|$ and the three point training set. Express the value of the discriminant function in the dual form for $x_2 = 0$. If we set $C < 1$, do we necessarily get a positive slack $(\xi_2 > 0)$ for this example $(x_2 = 0, y_2 = -1)$? Briefly justify your answer.

ξ_2 will be non-zero. The margin constraint for $(x_2 = 0, y_2 = -1)$ without slack requires that

$$\begin{aligned} y_2(\alpha_1 y_1 K(0,-1) + \alpha_2 y_2 K(0,0) + \alpha_3 y_3 K(0,1)) &= -1(\alpha_1 - \alpha_2 + \alpha_3) \\ &= \alpha_2 - \alpha_1 - \alpha_2 \\ &\geq 1 \end{aligned}$$

However, this constraint cannot be satisfied with $0 \leq \alpha_2 \leq C < 1$.

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) Will Algorithm 1 always converge (stop updating)? Do we need any additional conditions to ensure that it will?

If the points are distinct, the algorithm will always converge. This is because the radial basis kernel guarantees that the problem is linearly separable with a finite margin.

2.2 (6 points) Check all that apply

(X) a) If the algorithm converges, it finds a solution whose margin we can calculate. This margin depends on the order in which we cycle through the training examples

(X) 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.

(X) c) The number of mistakes that the algorithm makes (if it converges) depends on the value of σ^2

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

Answer 1: If σ^2 is small enough, then the perceptron algorithm will potentially make a mistake on each of the training examples. After these updates, all the points are correctly classified (there's no interference from other points).

Answer 2: As $\sigma^2 \to 0$, the margin we attain for each point is $1/\sqrt{n}$. Since $K(\underline{x}, \underline{x}) = 1$, the number of mistakes can be at most $1/\gamma_g^2 = n$.

Problem 3

A friend of ours claimed that she can reproduce 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 with maximum margin linear classifiers of the form

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

We have n labeled training examples: $\underline{\phi}(\underline{x}_1), \ldots, \underline{\phi}(\underline{x}_n)$ with ± 1 labels $y_1, \ldots, y_n$. You can assume that the training examples can be correctly classified with some $\underline{\theta}$.

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

$y_1\underline{\phi}(\underline{x}_1), ..., y_n\underline{\phi}(\underline{x}_n)$, i.e., we pass on the product of label and the feature vector as positive examples.

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?

$\hat{\underline{\theta}}$, 0, $\underline{\phi}(\underline{x})$

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 $\hat{\rho}$.

The margin is $\hat{\rho}/\|\hat{\underline{\theta}}\|$, i.e., the separating margin from origin to the "positive points" in the anomaly detection method.

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, $\hat{y} = \text{sign}\big(\, \underline{\theta} \cdot \underline{\phi}(\underline{x}) + \theta_0 \,\big)$, using just the two routines. Is this possible? Please justify your answer briefly.

No, but close. We can append the feature vectors with 1 such that $\underline{\phi}'(\underline{x}) = [\underline{\phi}(\underline{x}); 1]$ and $\underline{\theta}' = [\underline{\theta}; \theta_{d+1}]$. As a result, $\underline{\theta}' \cdot \underline{\phi}'(\underline{x}) = \underline{\theta} \cdot \underline{\phi}(\underline{x}) + \theta_{d+1}$ and we could proceed as before with the appended feature vectors. However, the anomaly detection method would minimize $\|\underline{\theta}'\|^2/2 = \|\underline{\theta}\|^2/2 + \theta_{d+1}^2/2$ and thus penalize larger values of the offset parameter θ_{d+1}. The answer would not be the same as maximum margin classifier with offset.

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 Task 1 and 2, respectively. The corresponding estimation problem is given by

$$\begin{aligned}
&\text{Minimize } \frac{1}{2}\|\underline{\theta}\|^2 \quad \text{with respect to } \underline{\theta},\ b_1, \text{ and } b_2, \text{ subject to} \\
&\quad \text{Task 1:} \quad -1(\underline{\theta} \cdot \underline{\phi}_i - b_1) \geq 1 \ \text{ if } y_i = 1, \quad +1(\underline{\theta} \cdot \underline{\phi}_i - b_1) \geq 1 \ \text{ if } y_i > 1 \\
&\quad \text{Task 2:} \quad -1(\underline{\theta} \cdot \underline{\phi}_i - b_2) \geq 1 \ \text{ if } y_i \leq 2, \quad +1(\underline{\theta} \cdot \underline{\phi}_i - b_2) \geq 1 \ \text{ if } y_i = 3 \\
&\text{for all } i = 1, \ldots, n
\end{aligned}$$

4.1 (3 points) Briefly explain why we would like to make sure that $b_1 \leq b_2$?

$b_1 \leq b_2$ ensures that the labels from the two binary classifiers are always consistent.

4.2 3If the problem is separable in the sense that the quadratic program has a solution, and all the rating labels occur at least once, are we guaranteed that the solution $\hat{\underline{\theta}}$, $\hat{b}_1$, $\hat{b}_2$ satisfies $\hat{b}_1 \leq \hat{b}_2$?

Y

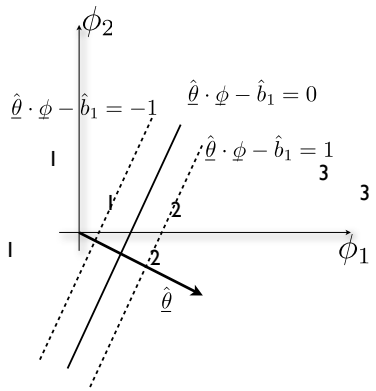

Figure 4.1. Student exam ratings in the feature space.

4.3 (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 approximately the resulting decision boundary and margin constraints in Figure 4.1 based on the data in the figure.

4.4 (6 points) How could you check if the solution $\hat{\underline{\theta}}$ based on Task 1 constraints alone also works as a solution to the combined task?

If we can find b_2 such that $\hat{\underline{\theta}}$ together with this b_2 satisfies Task 2 constraints, then it is the optimal solution to the combined task ($\|\hat{\underline{\theta}}\|^2$ cannot be smaller because of Task 1 constraints and Task 2 constraints are also satisfied).