Partial solutions to practice problems for Exam 1

COMS 4771 Fall 2016

Problem 1

The first and third ones.

Problem 2

In both (a) and (b), it is $n/(x_1 + \cdots + x_n)$.

Problem 3

(a) Yes.

(b) No.

(c) Let m_i be the number of mistakes in examples $1, \ldots, i-1$, and let T be the total number of mistakes. The number of kernel evaluations required to compute i-th prediction is m_i. So the total number of kernel evaluations is $\sum_{i=1}^n m_i$. The total number of mistakes overall is no more than 100. If the T-th mistake is made on the l-th example, then $m_{l+1} = m_{l+2} = \cdots = m_n = T$. In this case,

$$\sum_{i=1}^n m_i = \sum_{i=1}^l m_i + \sum_{j=l+1}^n T = nT - \sum_{i=1}^l (T - m_i),$$

where $m_i < T$ for $i = 1, 2, \ldots, l$. This is quantity is maximized when $l = T$, i.e., mistakes are made on the first T examples. In this case, we have $m_i = i - 1$ for $i = 1, 2, \ldots, T$, and hence

$$\sum_{i=1}^n m_i = nT - \sum_{i=1}^T (T - i + 1) = nT - \frac{T(T+1)}{2}.$$

If $n \leq 100$, then this quantity is at most $(n - 1/2)^2/2$. If $n > 100$, then this quantity is at most $100n - 5050$.

Problem 4

- ERM→A
- SVM→B
- Batch Perceptron→C
- Online Perceptron→D

Problem 5

(a) True.

(b) False.

Both Gini index and "classification error" are minimized when the examples reaching that leaf all have the same label.

(c) False.

The classifier $\hat{f}$ depends on both S and V (through ϕ), and hence is not necessarily independent of V.

(d) True.

The classifier $\hat{f}$ *is* independent of T.

(e) False.

No; the classifier has a quadratic decision boundary if the class conditional covariances are not equal.

(f) False.

This one was poorly worded. In lecture we showed how to obtain the linear classifier with the smallest training error rate on a set of examples S using a linear program, *when S is linearly separable.* This approach does not work in general.

(g) False.

The inner product between the weight vector $\hat{w}_t$ and $w_\star$ increases, but this could happen when the vector $\hat{w}_t$ gets longer, even if the angle between $\hat{w}_t$ and $w_\star$ actually increases.

(h) False.

No, consider kernel function $K(x, x') = \langle x, x' \rangle$.

(i) False.

You should try to construct a counterexample yourself!

(j) False.

No, soft-margin SVM always has a solution.