Assignment 4

Machine Learning COMS 4771

Spring 2017, Itsik Pe'er

Assigned: March 1st Due: Friday, March 10th, 1:10pm (note unusual deadline!)

Submission: Courseworks.

1) Multiclass SVM: When generalizing SVM to k>2, there are several options:
   a) One vs. rest: Apply pairwise SVM k times, where the pairwiseSVM$_i$ classifies class i vs. all the other classes. Show this idea is problematic by giving an example that results in points that the classifier cannot classify because it thinks they fit into multiple classes or none at all.
   b) One vs. one: Apply pairwise SVM k(k-1)/2 times, where the pairwiseSVM$_{ij}$ classifies class i vs. class j . When classifying a new point, we choose the class that wins the largest number of the pairwise comparisons. Show this idea is problematic by giving an example that results in points whose attempted classification is ambiguous because there are two more classes that draw for most pairwise wins.
   c) Optimizing all the margins together: The multiclass classifier includes an inequality above a hyperplane (in linear SVM) or curved surface (in kernelized SVM) per class, choosing to classify a point to the class whose inequality is satisfied the most, i.e. when presented as an inequality with zero produces the largest number. Instead of minimizing the reciprocal squared margin around the classifying inequality boundary as in pairwise SVM, this version of multiclass SVM minimizes the sum of reciprocal squared margins, across all k boundaries. Optimization is similarly subject to constraints of correct, margin-tolerant classification of training data. As in pairwise SVM, a single slack variable per datapoint is tolerated for the entire joint optimization problem, with the target function is penalized for the sum of these variables. Formally write the primal and quadratic problem for a general feature vector, and the dual one for a general kernel.

[50 points]

2) You assume the class assignment y_i for each item x_i you are classifying can receive values j=1,…,k with respective probabilities $p_1,\ldots,p_k$ .You are computing the log-likelihood for $Y=\{y_1,\ldots, y_N\}$.
   a) Based on your assumption, what is your expectation for the contribution of each y_i to the log-likelihood? Do you recognize the expression you got?
   b) While you didn't know that when computing the log-likelihood, your assumption was wrong. In fact, the respective probabilities are $q_1,\ldots,q_k$. Given that information, what is your revised expectation for the contribution of each y_i to the log-likelihood (that you computed without knowing this information)?

[30 points]

3) Ensemble:

a) When Bagging n elements out of n with replacement, what is the marginal distribution of each bootstrap weight? What is the correlation of each two?

b) You are using T independent weak learners that each guarantee $\frac{1}{2} - \gamma$ worst case error. You want a majority-vote booster that would guarantee ϵ worst-case error. Show that there exists some C that is independent of γ, ϵ such that $T = -\frac{C \log \epsilon}{\gamma^2}$ would suffice

[30 points]

Good luck!