Assignments for TIML-10 (CS-689)

Note: The marks, deadline and difficulty for each of the problems are marked. A problem with a '$*$' may require use of results obtained by others which can be got through internet/library etc. These problems are not expected to be solved by the students without such aids. Problems marked '$**$' may require numerical computation/plotting. In general, problems not marked with stars MUST be solved by the student by himself.

1 Statistical Learning Theory

1. Suppose the optimal (linear) classifier for some 2-D data is $\text{sign}(\mathbf{w}^\top \mathbf{x} - b = 0)$ where $\mathbf{w} = [0.6667,\ -1]^\top, b = 0$. Suppose a test data point $X = [X_1,\ X_2]^\top$ is given; however X is not known exactly (it is uncertain, say modeled by a random variable). What is known about X is that it has mean $[2,\ 0.75]^\top$, $0.5 \leq X_1 \leq 2.5$, $0.5 \leq X_2 \leq 1.5$ and it can be assumed that X_1 is independent of X_2. Can you now lower-bound the probability that the given test data point lies on the positive side of the line $\mathbf{w}^\top x - b = 0$? (Hint: Use the Hoeffding and Chernoff bounds introduced in lecture-1).

   [1 Mark, 26-Jan-09]

2. Derive some kind of uniform convergence based sufficient conditions for consistency of *Empirical Risk Maximization*. Your proofs must not use the results for the case of ERM (minimization).

   [1 Mark, 26-Jan-09]

3. Prove that the VC dimension of hyperplane classifiers (linear discriminators) in $\mathbb{R}^d$ is $d + 1$.

[1 Mark, 26-Jan-09, $*$]

4. Suppose $\mathcal{F}$ is the set of all functions f which are "elliptical discriminants" in $\mathbb{R}^2$ and are parallel to the x and y axes i.e., $f(x) = \text{sign}\left((x-c)^\top S(x-c) - 1\right)$ where $c \in \mathbb{R}^2$ and S is a 2×2 diagonal matrix with non-negative entries. What is the VC dimension of $\mathcal{F}$? Now, suppose $\mathcal{F}$ additionally also includes elliptical discriminants whose principal axes are at 45 degrees to the x and y axes. What is the VC dimension of this modified $\mathcal{F}$? Give (intuitive) justifications for your answers.

[1 Mark, 26-Jan-09]

5. Based on above problem you can now imagine what would be the definition of polynomial discriminators (else google!). What is the VC dimension of such an m degree polynomial discriminator in n dimensional space ?

[1 Mark, 26-Jan-09, $*$]

6. In class we derived an upper bound on the probability that the worst-case difference between true and empirical risks is greater than a tolerance ϵ involving VC dim. Now, plot (either print or hand-drawing) how this bound varies with ϵ and m (the no. examples) for a hyperplane classifier in 2-d. For what m and/or ϵ is this bound useless ?

[1 Mark, 26-Jan-09, $**$]

2 Support Vector Machines

1. Consider two variants of the soft-margin SVM formulation in the input-space (linear kernel), where the squared Euclidean-norm (l_2-norm) regularizer is replaced with l_1-norm (i.e., sum of absolute values) and l_∞-norm (i.e., maximum of absolute values). You may want

to pose these new variants as Linear Programs (LPs) or even otherwise using MATLAB (optimization toolbox) or SeDuMi[1] or Mosek[2] or cvx[3] or any other opt. toolbox[4] you are familiar with, solve these three formulations. You might also want to use your favorite SVM solver[5] for tackling the usual l_2-norm formulation. Generate synthetic data, in say 10-d (consider linearly separable data as well as almost linearly separable data), and compare the optimal $\mathbf{w}$ obtained with the three formulations as a function of the regularization parameter C (consider $C = 0, 1^{-3}, 1^{-2}, \ldots, 10^2, 10^3, \infty$) on the synthetic data. Summarize your findings. Students might be asked for individual demos[6].

[5 Marks, 22-Feb-10, $**$]

2. In the lecture, we came up with some threshold values of W (the upper bound on $||\mathbf{w}||$) for which the VC-dimension of canonical hyperplane classifiers in $\mathbb{R}^2$ changes. Repeat the same in $\mathbb{R}^3$.

[1 Mark, 06-Mar-10]

3. Derive the dual of the (linear) SVM formulation using the square-hinge-loss rather than the hinge-loss. The dual resembles some formulation which we discussed in class; can you recognize it. Comment on the uniqueness of the dual solution. How do you recover $\mathbf{w}, b$ from the dual solution ?

[1 Mark, 06-Mar-10]

4. Consider the following kernel defined on Euclidean space:

$$k(x, y) = \begin{cases} 0 & \text{if } x = y, \\ x^\top y & \text{if } x \neq y \end{cases}$$

[1] http://sedumi.ie.lehigh.edu/

[2] Student's free trial at http://www.mosek.com/index.php?id=7

[3] http://www.stanford.edu/~boyd/cvx/

[4] Look at http://en.wikipedia.org/wiki/Linear_programming#Solvers_and_scripting_.28programming.29_languages

[5] http://www.support-vector-machines.org/SVM_soft.html

[6] With such a wide choice of re-formulations, solvers, data generation there is very less probability that two students give "same" answers :)

Is this a positive/negative/indefinite[7] kernel ? Justify your answer.

[1 Mark, 06-Mar-10]

5. Let Ω be a non-empty set of finite cardinality. Consider a kernel defined on the subsets of Ω i.e., $k : 2^{\Omega} \times 2^{\Omega} \mapsto \mathbb{R}$ given by[8]: $k(A_1, A_2) = \exp\{3|A_1 \cap A_2|^2\}$. Show that k is a positive kernel. You MUST give **two** "different" arguments/proofs for showing positiveness (may be come up with two "different" inner-product spaces where the given kernel or a related kernel is an inner-product).

[1.5 Marks, 06-Mar-10]

6. Describe the ν-SVM formulation highlighting its merits over usual SVM formulation.

[1 Mark, 06-Mar-10, *]

7. Consider the regression problem of predicting a real-world variable such as rainfall in a particular week. Assume that the training datapoints lie in a sphere of radius R in a Euclidean space $\mathbb{R}^d$ and $\mathcal{F}$ is set of all linear regressors in that space through origin with $\|\mathbf{w}\| \leq W$. Assume loss function is $l(f(x), y) = (y - f(x))^2$. Now justify the Ridge-regression formulation using SLT.

[1 Mark, 06-Mar-10]

8. Write a brief note on Martingale difference sequences and applications.

[1 Mark, 06-Mar-10, *]

9. Suppose the optimal discriminant function employed for a binary classification application with 2-d data is $f(X_1, X_2) = \text{sign}\{1.5(X_1-1)^2 + 1.5(X_2-1)^2 + (X_1-1)(X_2-1) - 1\}$, where X_1, X_2 are the two feature values of a datapoint X. You are now required to lower bound

[7] k is negative if $-k$ is positive. k is indefinite if it is neither positive nor negative
[8] Here, $|A|$ denotes cardinality of A.

probability that an uncertain (noisy) test datapoint X is labeled positive by this f. Only partial information regarding X is known: $-2 \leq X_1 \leq 2, -1 \leq X_2 \leq 1, \mathbb{E}[X_1] = 0.5, \mathbb{E}[X_2] = -0.5, var(X_1) = 0.75, var(X_2) = 0.25$ and X_1, X_2 are independent. Compute this upper bound using two different conc. ineq. discussed in class.

[1.5 Marks, 06-Mar-10]

3 Kernel Learning

1. Solve the MKL formulation by posing it as a QCLP (Quadratically Constrained Linear Program). Again, you are free to use your favourite solver. Run your solver on random 50 datapoints sampled[9] from the Sonar dataset[10] taking five base kernels which are all Gaussian kernels but with different parameters: $10^{-2}, 10^{-1} \ldots, 10^{2}$ and $C = 100, B = 1$. How do you obtain optimal kernel weights from the solution ? Are the kernel weights always uniquely determined? Are they uniquely determined in this case ? If yes, can you play around with the choice of your kernels so that the soln. is not unique ? If no, does the choice lead to differences in terms of test accuracy ? Verify that your code is correct by comparing it with reduced-gradient based solver[11]. Is the problem equivalent to solving a single SVM with the "optimal" kernel ("optimal" kernel is any base kernel that has non-zero kernel-weight at optimality) ? If so, can you solve the MKL problem by simply running SVM code n (number of base kernels) times ? How ? How is MKL different than tuning kernel parameters using cross-validation ? How does MKL compare to an "SVM that uses kernel which is a simple sum of all the base kernels" in terms of accuracy (u may want to tune C parameter for MKL and SVM for doing this comparison)? Can u come-up with situations where MKL will beat SVM convincingly and vice-versa? As far as possible justify your answers through the simulation results.

[9] the remaining can be used as validation/test sets

[10] http://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+(Sonar,+Mines+vs.+Rocks)

[11] Available at http://asi.insa-rouen.fr/enseignants/~arakotom/code/mklindex.html

[10 Marks, 19-Mar-10, **]