Exercise 1: Active Learning

Distributed 2.4.20, Submission 23.4.20

1 Testing a simple active algorithm

The aim of this exercise is to introduce you to some basic features of Active Learning (AL) through the implementation and study of a simple algorithm within a pool based active learning setting.

The knowledge required for this exercise is:

1. The basic elements of supervised learning.
2. Applying the linear and nonlinear Support Vector Machine (SVM) algorithm. Use any reputable code for this. For the nonlinear SVM we use the RBF kernel.
3. Error estimation through Leave-One-Out (LOO) cross validation.

Preparation

1. Read sections $1-4$ of the paper "Support Vector Machine Active Learning with Applications to Text Classification", by Tong and Koller.
2. Write a program to generate a synthetic binary classification problem where the class labels assume the values $y = \pm 1$. The conditional distribution for each class is normal, namely $p(x|y=1) = \mathcal{N}(x; \mu_+, \Sigma_+)$, where μ_+ and Σ_+ are the mean and covariance of the random vector $x \in \mathbb{R}^d$, conditioned on $y = 1$, and similarly for $y = -1$. The prior class probabilities are equal, $p(y = \pm 1) = 1/2$.
3. Download the two medical based datasets `breast-cancer` and `diabetes` from the LIBSVM data repository, where you can find information about their properties.
4. For all datasets in this exercise set aside 300 randomly chosen points for training (the pool), 150 for validation and 150 for testing. These sets should, of course, be disjoint. The test set should be used only for assessing performance.

5. For the RBF classifiers use the a kernel of the form $k(x, y) = \exp\left(-\gamma\|x - y\|^2\right)$. Use the validation set to determine the values of the parameter C (for both linear and RBF kernels) and γ for the RBF kernel. Both these parameters *must* be chosen based on the validation set and not set a-priori. All reasonable packages contain a method to do this. Do this for each data set in advance, and use the estimated parameters in all experiments on that dataset. Even if you used a software package to do this, explain how it is done.

We will consider two basic algorithms, the first of which yields the standard passive learning setup.

1. Rand - Query a point chosen uniformly at random from the unlabeled set and request its label.
2. Simple - Query a point that is closest to the decision boundary of the SVM. Use the algorithm proposed in the paper *Support vector machine active learning with applications to text classification*, Tong and Koller, JMLR 2001.

Tasks

1. Synthetic data:
   (a) Consider the case $\Sigma_+ = \Sigma_-$. Show that a linear classifier $\hat{y} = w^\top x + b$ is optimal in this setting and compute w and b in terms of $\mu_\pm$ and Σ.
   (b) What is the optimal classifier form when $\Sigma_+ \neq \Sigma_-$?
   (c) Choose $\mu_+ = (1, 1, \ldots, 1)/\sqrt{d}$ and $\mu_- = -(1, 1, \ldots, 1)/\sqrt{d}$ and $\Sigma_+ = \Sigma_- = I$, the unit matrix in d dimensions. Estimate the Bayes error for $d = 1, 20, 50, 200$. Repeat these results for $\mu_\pm = \pm(1, 1, \ldots, 1)$. Explain the differences in the Bayes error.

Repeat the following experiments for the three datasets (two natural and one artificial) and each of the classifiers (linear and RBF SVMs). The unlabeled pool is taken to consist of the set of 300 data points set aside for training. Each experiment should be performed 30 times, and the average results should be displayed. For experiment repetition choose the pool/validation/test sets randomly (so as to obtain independent experiments).

2. For each dataset, select an initial set of 5 points out of the training set, label them, and construct an initial classifier. Run the algorithm by querying $n =$

$20, 40, \ldots, 280$ additional points from the training set. For each n plot the test error as a function of the number of queried labels (test error is calculated using the designated test set).

3. Compute the LOO cross validation estimate for each case (dataset/algorithm) and compare to the test error calculated in the previous item. Discuss the quality of the estimate for each case, and explain what you observe. Note that LOO estimates the error based on the n labeled points.

4. Repeat the above for the first synthetic Gaussian problem above, with dimensions $d = 2, 50, 100, 200$ and discuss your results. Relate your results to the Bayes error.

5. Can you think of a dataset for which the performance of Simple is expected to be poor? Create such a dataset and test your hypothesis.

2 Theoretical exercises

I VC dimension

Compute the VC-dimension of the following sets. Sets correspond to classifiers by labeling each point in the set as $+1$ and the complement as -1.

1. Subsets of the real line formed by the union of k intervals?

2. The set of subsets of the real line parameterized by a single parameter α : $I_\alpha = [\alpha, \alpha + 1] \cup [\alpha + 2, \infty]$.

3. Right triangles in the plane with the sides adjacent to the right angle both parallel to the axes and with the right angle in the lower left corner.

II From probabilities to expectations

Assume a non-negative random variable Xobeys the inequality $P[X > t] \leq ce^{-2nt^2}$for all $t > 0$ and some $c > 0$. Show that

$$E\left[X^2\right] \leq \frac{\log(ce)}{2n} \qquad ; \qquad E[X] \leq \sqrt{\frac{\log(ce)}{2n}}\,.$$

III Disagreement

Let $\mathcal{H}$ be a space of binary hypotheses over $\mathcal{X}$ and D a distribution over $\mathcal{X}$. Define a metric over $h \in \mathcal{H}$, and a ball in hypothesis space,

$$d\left(h, h'\right) = P\left\{h(X) \neq h'(X)\right\} \qquad ; \qquad B(h, r) = \left\{h' \in \mathcal{H} \,:\, d(h, h') \leq r\right\} \,. \tag{1}$$

For any $V \subseteq \mathcal{H}$ define the disagreement region

$$\mathrm{DIS}(V) = \left\{x \in \mathcal{X} \,:\, \exists h, h' \in V \,\text{with}\, h(x) \neq h'(x)\right\} \,. \tag{2}$$

1. Prove that for any distribution D, $d(h, h')$ is a distance measure.
2. Let C be the class of linear classifiers through the origin in $\mathbb{R}^d$, and let D be the uniform distribution over the unit ball in $\mathbb{R}^d$. Let h be the linear classifier $h(x) = I\left(x_1 \geq 0\right)$. What is DIS($B\left(h, \epsilon/\sqrt{d}\right)$)?