Assignment 3: Hidden Markov Models, Agnostic PAC Learning and VC dimension

CS485/685 – Winter 2012

Out: March 1, 2012
Due: March 15, 2012, at the beginning of the lecture. Late assignments may be submitted in the pink drop off box on the third floor of MC within 24 hrs for 50% credit.

Be sure to include your name and student number with your assignment.

1. **[40 pts]** Let X be some domain set and H a collection of functions from X to $\{0,1\}$. For a probability distribution D over $X \times \{0,1\}$ let $L_D(H)$ denote the minimum true loss $inf\{L_D(h) : h \in H\}$ and, for sample $S \subset X \times \{0,1\}$, let $L_S(H)$ denote the minimum empirical loss $min\{L_S(h) : h \in H\}$. We say that H is *ERM-learnable* (where ERM stands for Empirical Risk Minimization), if for every positive ϵ and δ there exist a function $m(\epsilon, \delta)$ so that, for every probability distribution D over $X \times \{0,1\}$, if S is an i.i.d. sample from D of size $N \geq m(\epsilon, \delta)$, then

$$P_{S \sim D^N}[sup\{|L_D(h) - L_D(H)| : h \in H \text{ and } L_S(h) = L_S(H)\} > \epsilon] < \delta$$

(That is, with high probability over the choice of S, an h that minimizes the S-empirical loss has true loss that is close to that of the best hypothesis in H).

   (a) **[10 pts]** Let X be the 12-dimensional Euclidean space $\Re^{12}$ and define a function h by

$$h(\bar{x} = (x_1, \ldots, x_{12})) = 1 \text{ if } \sum_{i=1}^{12} \frac{x_i}{2^i} > 0.5$$

   and $h(\bar{x}) = 0$ otherwise. Find a sample size, N_0, such that for any probability distribution D over $\Re^{12} \times \{0,1\}$ for which $L_D(h) = 0.3$, if S is an N-size sample drawn i.i.d by D, then with probability greater than 0.9, $0.2 \leq L_S(h) \leq 0.4$. The smaller the sample size N_0 you come up with, the higher your mark will be. However, you should prove your claim.

   (b) **[20 pts]** Let $X = [0,1]$, let

$$H_{mirrow} = \{h : X \to \{0,1\} : \forall x h(x) = 1 - h(1-x)\}$$

   Prove that H_{mirror} is *not* ERM-learnable. Hint: for D, consider the uniform distribution over $[0, 0.5] \times \{1\}$.

   (c) **[10 pts]** Given a class of functions, H, as above, define a new class $H^C = \{h^c : h \in H\}$ where, for $h : X \to \{0,1\}$, the function h^c is defined by setting, for every x, $h^c(x) = 1 - h(x)$. Prove that a class H is ERM learnable if and only if the class H^C is ERM learnable.

2. **[10 pts]** What is the VC-dimension of the class of all disks in the Euclidenan plane? (By a disk we mean a set of the form

$$d_{x_0, y_0, r} = \{(x, y) : (x - x_0)^2 + (y - y_0)^2 \leq r\}$$

for some $(x_0, y_0) \in \Re^2$ and $r \in \Re$). Prove your claim.

3. **[50 pts]** Hidden Markov Models

In this question, you will experiment with a Hidden Markov Model (HMM) and compare it to the mixture of Gaussians model that you implemented in Assignment 1. Download the dataset posted on the course website. It consists of several sequences of continuous inputs and discrete outputs. The goal is to infer the outputs based on the inputs. As a baseline, train a mixture of Gaussians model and classify each instance separately. Then train a hidden Markov Model and classify the instances by taking into account the correlations between them.

(a) **[10 pts]** Train a hidden Markov Model by supervised maximum likelihood learning with the training set. Since the inputs are continuous, estimate the mean and covariance matrix of the Gaussian emission distributions. Since the outputs are discrete, estimate the parameters of the multinomial transition distributions and initial state distribution. Similarly, train a mixture of Gaussians model.

What to hand in:

- Printout of your code.
- Printout of the parameters of the HMM and Mixture of Gaussians.

(b) **[20 pts]** Implement the Forward algorithm for monitoring with your HMM. More precisely, estimate the probability of the class at each step based on the current and previous inputs by computing $\Pr(Y_t|X_{1..t})$. In comparison, estimate the probability of the class at each step based on the current inut only with the mixture of Gaussian models by computing $Pr(Y_t|X_t)$. For both models, return the class that has the highest probability at each step.

What to hand in:

- Printout of your code.
- Monitoring accuracy (percentage of correctly classified instances in the test set) of the HMM and Mixture of Gaussian models.
- Discuss the results.

(c) **[20 pts]** Implement the Viterbi algorithm for simultaneous classification of all the instances with the HMM. More precisely, find the sequence of outputs that has the highest probability given all the inputs by computing $argmax_{Y_{0..t}} \Pr(Y_{0..t}|X_{1..t})$.

What to hand in:

- Printout of your code.
- Joint classification accuracy (percentage of correctly classified instances in the test set).
- Discuss and compare the results found for monitoring and joint classification.