

ASSIGNMENT 1, due Thurs, February 1

Discussions are encouraged but no written records can be taken away from a discussion. Books and notes can be consulted but not copied from. Homeworks are due in class or in my mail box by 4pm on the due date.

1 (50 pts) Learning a Language

Discuss what kind of learning task learning a language would fall under. (Remember that in class we looked at three types of learning tasks – Supervised, Reinforcement and Unsupervised).

2 (50 pts) Probability

There are 2 black opaque bags, each containing 2 balls. One bag has 2 black balls and the other has a black and a white ball. You pick a bag at random and then pick one of the balls in that bag at random. When you look at the ball it is black. You now pick the second ball from that same bag. What is the probability that this ball is also black?

3 (200pts) Computation of the Test Error

For the example used in class, the data set was $\mathcal{D}_N \equiv \{(0, 0, 0, 0), (0, 0, 1, 1), (0, 1, 0, 1), (0, 1, 1, 0)\}$. The fourth entry in the vector is the y value. Remember that the target function was given by $y = x_1 \oplus x_2 \oplus x_3$. The measure of performance was the probability of error ($P[err]$) on a randomly chosen input point (each input point has a probability $1/8$ of being selected). In class we considered a learning model (\mathcal{L}) that contained only the 4 functions $\{AND, OR, NAND, NOR\}$. Now, let the learning model be the entire set of boolean functions on 3 boolean input variables. There are 2^{2^3} such functions. The learning algorithm will be as follows. Find all functions that attain a training $P[err]$ of ν and pick one of them randomly with equal probability. In this problem we will analyse the ultimate performance.

(a) Show that

$$Test P[err] = P[err|x \in D] P[x \in D] + P[err|x \notin D] P[x \notin D]$$

Hence show that for this learning system, $P[err] = N\nu/8 + (8 - N)/16$, where N is the number of data points, in this case 4. Notice that the test error is a monotonic function of the training error. Why is this good?

- (b) The Off Training Set (OTS) error is the probability of error on a randomly selected point that is not in the training set. Assume that $N < 8$. Show that $P_{OTS}[err] = 1/2$, independent of the data set, or target function!
- (c) In this example we know the target function. Suppose instead that you did not know the target function, but that I told you that the target function was symmetric with respect to its arguments (a function is symmetric with respect to its arguments if the function remains the same when you permute the arguments). Suggest ways in which the situation in part (b) could be improved. (hint: the only things at your disposal are the learning algorithm and the learning model)
- (d) Why might OTS error be a more relevant error measure than the test error?

4 (200pts) No Free Lunch (NFL)

For this problem assume that the input space is finite and one dimensional. Once again, the performance measure we use is the probability of error. For concreteness let the input space be $\{1, 2, \dots, M\}$. The data set consists of some set of N points, together with the value of the (boolean) function at those points.

- (a) Suppose that the target function was generated as follows: some super-power picked a function at random, with each function having an equal probability of being chosen. Thus, every function consistent with the data set is equally likely to be the target function (f). Let ν be the training error for ANY function g . For this function g , show that the test error is given by

$$P[err] = \frac{N\nu}{M} + \frac{M - N}{2M}$$

where the probability in this case is with respect to the random choice of functions.

- (b) What is the OTS error? Show that in the limit $M \rightarrow \infty$, the OTS error and the test error converge to the same value. What is the test error in the limit $N \rightarrow M$?
- (c) Prove the following No Free Lunch theorem (for this learning scenario):

Theorem: *For any learning algorithm and any learning model, if the probability of any given target function (consistent with the data set) is uniform then the OTS error is a constant, independent of the details of the learning system. The same result holds in the asymptotic limit $M \rightarrow \infty$ for the test error as well.*

(in other words, no matter what we do, we end up with the same OTS error. In particular the algorithm that picks the function with the largest training error in the learning model and the algorithm which picks the function with smallest training error perform equally well on unseen data points).

- (d) The situation in (c) seems dire. It appears that no amount of “learning” can help. Does this mean we should give up? Why not?

5 (150pts) The Ubiquitous Gaussian - Central Limit Theorem

In this problem we will study why the Gaussian assumption is not such a bad assumption. Usually a random fluctuation of values is caused by lots of little fluctuations adding up (or averaging). These little fluctuations are independent and can be from different distributions. The central limit theorem says that when you add up lots of little independent fluctuations, what you end up with is nearly Gaussian. We will demonstrate this with the following surprising experiment where the total fluctuation is the sum of (only) 10 little fluctuations.

- (a) Write a program that will generate independently 10 random numbers. The i^{th} number is generated uniformly from 1 to $2 - 1/i$. For example the first is from 0 to 1, the second from 0 to $1\frac{1}{2}$, the third from 0 to $1\frac{2}{3}$ etc. Now take the average of these numbers.
- (b) Repeat (a) 1000000 times to get 1000000 averages. Compute the mean and variance of these numbers.
- (c) Plot a histogram of these numbers and on the same plot show the Gaussian with the same mean and variance as the 1000000 numbers. The two plots should be nearly identical. What you see should be surprising. Even though each little fluctuation is drawn from a *different* uniform distribution, the overall fluctuation is very nearly Gaussian. This is the reason that we usually assume Gaussian densities. [Note: you need to scale by the bin width \times the number of samples to convert a density into a histogram.]

6 (50pts) Minimum error rate classification

Let the states of the world be $w_1 \dots w_S$. Associated with each state is the action α_i , $i = 1 \dots S$. One can view the α_i as a prediction of the state of the world. This problem is usually referred to as classification and if one takes action α_i , one says that one has predicted class i . For example if one takes action α_1 then one predicts class 1. Suppose that the loss matrix is given by

$$\lambda_{ij} = \begin{cases} 1 & i \neq j \\ 0 & i = j \end{cases}$$

- (a) Show that the conditional risk $R(\alpha_i|x)$ is equal to the probability of error (i.e., the probability of predicting the state incorrectly).

7 (100pts) Construction of a Bayes' Optimal Rule

In this problem, the feature vector (x) has two (binary) components, (x_1, x_2) . There are 2 states of the world w_1 & w_2 and there are two actions α_1 & α_2 . The joint probability distribution $P(x, w)$ and the loss matrix (Λ) are given by

$$P(x, w) : \begin{array}{l} x = (0, 0) \\ x = (0, 1) \\ x = (1, 0) \\ x = (1, 1) \end{array} \begin{array}{c} w_1 \quad w_2 \\ \left(\begin{array}{cc} 0.25 & 0.05 \\ 0.05 & 0.05 \\ 0.2 & 0.1 \\ 0.1 & 0.2 \end{array} \right) \end{array} \quad \Lambda : \begin{array}{c} w_1 \quad w_2 \\ \alpha_1 \left(\begin{array}{cc} 1 & 10 \\ 5 & 0 \end{array} \right) \\ \alpha_2 \end{array}$$

Compute the optimal Bayes' decision function. Remember that a decision function is a mapping from the input space to the possible actions.

8 (200pts) Discriminant Function for the Normal Density

Suppose that the feature vector is d -dimensional and that $P(x|w_i) \sim N(\mu_i, \Sigma_i)$. In class we assumed that $\Sigma_i = I$ for all i . In this problem, assume that the covariance matrix is the same for each i , but not necessarily the identity matrix. Namely, $\Sigma_i = \Sigma$, a constant matrix. The Normal density is given by

$$P(\mathbf{x}|w_i) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (\mathbf{x} - \mu_i)^T \Sigma^{-1} (\mathbf{x} - \mu_i) \right]$$

- (a) Compute the Bayes' optimal minimum error rate discriminant functions (i.e., the loss matrix is the 0-1 loss matrix as in problem 6).
- (b) Specialize to the 2-state/2-actions case. Relabeling $\alpha_1 = 1$ and $\alpha_2 = -1$, show that the optimal action can be written in the form

$$\alpha = \text{sign}(w^T x + w_0)$$

and compute w, w_0 . (Remember that in the two actions case, a single discriminant function which is compared to zero suffices).

9 Time for Homework

- (a) How long did this problem set take to do?
- (b) Is the pace of the class too fast/slow? Answer on a scale 1(slow) to 10(fast).

Please answer Problem 8 on a separate sheet of paper and hand it in separately, WITHOUT your name on it.