

## ASSIGNMENT 1 Solutions

### 1 (50 pts) Learning a Language

Could be any combination of supervised, unsupervised, reinforcement:

Supervised: A teacher is providing direct feedback on what the correct way to say things is.

Unsupervised: Eg. Watch TV, read a lot, hear native speakers then become better at the language.

Reinforcement: For example say something asking for water and you either get water (positive feedback) or not (negative feedback).

### 2 (50 pts) Probability

Let  $B$  be black ball and  $W$  be white ball. Subscripts indicate whether we are talking about the first or second ball.

$$P[B_2|B_1] = \frac{P[B_2 \text{ and } B_1]}{P[B_1]} = \frac{1/2}{3/4} = \frac{2}{3}$$

### 3 (200pts) Computation of the Test Error

(a) Since  $x$  is either in  $D$  or not.

$$\text{Test } P[\text{err}] = P[\text{err and } x \in D] + P[\text{err and } x \notin D]$$

Using Bayes theorem now gives the result want.

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

$P[\text{err}|x \in D]$  is just  $\nu$  and  $P[x \in D]$  is  $N/8$ .  $P[x \notin D]$  is  $(8 - N)/8$  and  $P[\text{err}|x \notin D] = 1/2$  because half the functions are 0 and half the functions are 1 on every point outside the training set and so since  $f(x)$  is either 0 or 1, the probability of error is 1/2.

Monotonic is good because if you pick lowest training error, you get lowest test error.

(b)  $P_{OTS}[\text{err}]$  is exactly  $P[\text{err}|x \notin D] = 1/2$  as discussed above.

(c) Restrict the set of functions to only symmetric functions or equivalently of the functions with given  $\nu$  only pick ones that are symmetric.

(d) Because performance on the training set is memorization and one is usually only interested in performance on un seen data.

## 4 (200pts) No Free Lunch (NFL)

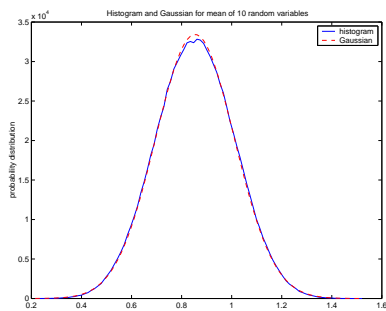
- (a) Since  $P[x \in D] = N/M$  and  $P[x \notin D] = (M-N)/M$ , using 3(a) we get  $P[err] = \frac{N\nu}{M} + \frac{M-N}{M}P[err|x \notin D]$  and since  $f(x) = 0$  with probability  $1/2$  on every  $x$ ,  $P[err|x \notin D] = \frac{1}{2}$  yielding the desired result. [Note that the target function  $f(x)$  plays the role of the  $g$ 's in the previous problem and  $g(x)$  plays the role of  $f$ .
- (b)  $P_{OTS} = P[err|x \notin D] = \frac{1}{2}$ . From part (a),  $P[err] = \frac{1}{2} + N(\nu - \frac{1}{2})/M$  and as  $M \rightarrow \infty$ , the second term vanishes. Thus  $\lim_{M \rightarrow \infty} P[err] = P_{OTS}[err] = \frac{1}{2}$ . As  $N \rightarrow M$ , the second term becomes  $\nu - \frac{1}{2}$  and so we get  $\lim_{M \rightarrow N} P[err] = \nu$ .
- (c) **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.

PROOF  $P_{OTS} = 1/2$  for every  $g(x)$  hence for the particular  $g(x)$  output by any learning algorithm. ■

- (d) Don't give up as there is still more to the class. The key is that target functions are not random and usually we know something about the function such as symmetry as in the example in problem 3.

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

- (a) see code on web
- (b) mean  $\approx 0.8525$  and variance  $\approx 0.0248$



- (c)

## 6 (50pts) Minimum error rate classification

- (a)

$$R(\alpha_i|x) = \sum_{j=1}^N \lambda_{ij} P[w_j|x] = \sum_{j \neq i} P[w_j|x] = 1 - P[w_i|x] = \text{prob of error}$$

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

$R(\alpha_1|x) = P(w_1|x) + 10P(w_2|x)$ ,  $R(\alpha_2|x) = 5P(w_1|x)$  so choose  $\alpha_1$  if  $R(\alpha_1|x) \leq R(\alpha_2|x)$ , i.e., if

$$\frac{P(w_2|x)}{P(w_1|x)} = \frac{P(w_2|x)p(x)}{P(w_1|x)p(x)} = \frac{P(w_2, x)}{P(w_1, x)} < \frac{2}{5}$$

$\frac{P(w_2, x)}{P(w_1, x)} < \frac{2}{5}$  only for  $x = x_1$ , therefore  $\alpha(x_1) = \alpha_1$ ,  $\alpha(x_2) = \alpha(x_3) = \alpha(x_4) = \alpha_2$ .

## 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) As discriminant functions we can use  $R(\alpha_i|x) = 1 - P(w_i|x)$  from problem 6. Minimizing  $R(\alpha_i|x)$  means maximizing  $P(w_i|x)$ , so we can use  $P(w_i|x)$  as discriminant function. But  $P(w_i|x) = P(x|w_i)P(w_i)/P(x)$ , so by taking the log and ignoring terms that do not depend on  $i$ , we get

$$\max_i \left\{ -\frac{1}{2} (\mathbf{x} - \mu_i)^T \Sigma^{-1} (\mathbf{x} - \mu_i) + \log P(w_i) \right\}$$

Further expanding and discarding  $-\frac{1}{2} \mathbf{x}^T \Sigma^{-1} \mathbf{x}$  we get

$$g(\alpha_i|x) = \frac{1}{2} \mu_i^T (\Sigma^{-1} + (\Sigma^{-1})^T) \mathbf{x} + \log P(w_i) - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i$$

where we use  $g(\alpha_i|x)$  to represent the discriminant function.

- (b)  $\alpha = \text{sign}(g(\alpha_1|x) - g(\alpha_2|x))$  gives the correct decision function. Substituting from above, we get

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

with

$$w = \frac{1}{2} (\Sigma^{-1} + (\Sigma^{-1})^T) (\mu_1 - \mu_2) \quad w_0 = \log \frac{P(w_1)}{P(w_2)} + \frac{1}{2} (\mu_2^T \Sigma^{-1} \mu_2 - \mu_1^T \Sigma^{-1} \mu_1)$$