Homework #5, EE556

Due: 10/30/18; Hand in problems 1,2, 3, and 6

Problem #1

Recall the "committee machine" developed in lecture, which simply forms the arithmetic average of a set of regression functions, *i.e.*, $f_{\text{com}}(\underline{x}) = \frac{1}{M}\sum_{i=1}^{M} f_i(\underline{x})$. In this problem, we extend this to the "weighted" committee machine, '*i.e.* $f_{\text{com}}(\underline{x}) = \sum_{i=1}^{M} w_i f_i(\underline{x})$, with the constraint that $\sum_{i=1}^{M} w_i = 1$. Considering a training set of T input-output examples, use the method of Lagrange multipliers to find the set of weights $\{w_i\}$ that minimizes the sum-of-squared regression errors while satisfying the constraint on the weights.

Problem #2

Problem 5.9, Haykin.

Problem #3

Show that a normalized Gaussian radial basis function classifier is equivalent to a Bayes classifier if the (feature-vector,class label) pairs are modeled by a mixture model with the following stochastic generation: 1) randomly select one of the basis functions (components); 2) given the component, choose the feature vector according to the Gaussian component density; 3) choose the class label according to a probability mass function that is conditioned on the component.

Problem #4

Again consider the normalized Gaussian radial basis function network, but this time for statistical regression rather than for classification. Justify the interpretation of this network as a conditional mean (minimum mean-squared error) estimator. Your justification should include specification of the associated statistical model (stochastic generation) for Y given $\underline{X} = \underline{x}$.

Problem #5

Explain why, for the case of constant width Gaussian basis functions, the basis function is defined *without* the term $\frac{1}{(2\pi\sigma^2)^{d/2}}$ – *i.e.*, without the term needed to make the basis function a true density function.

Problem #6

In class, we introduced the cross entropy "distance" between probability mass functions. In this problem we will elaborate on an application of this quantity to statistical pattern recognition. Frequently, in pattern recognition (and in other domains), one may have access to multiple, *different* estimates of a probability mass function. Each such estimator can be viewed as an "expert". Each expert may use a different probability model, or may be based on a different (training) data set. The problem is then how to combine these probability estimates to obtain an overall estimate. Consider J "expert" pmfs $\{P_j[k|x], k = 1, \ldots, K\}, j = 1, \ldots, J$.

i) Find the "combined" pmf that minimizes the average cross entropy cost to the experts:

$$\min_{\{P[k|x]\}} \frac{1}{J} \sum_{j=1}^{J} D(\{P[k|x]\}||\{P_j[k|x]\}).$$

Note: This is (actually) a constrained minimization problem that can be solved by the method of Lagrange multipliers. Q: What's the constraint ?

ii) Since cross entropy is an asymmetric cost, this time instead choose the combined probabilities to minimize:

$$\frac{1}{J} \sum_{j=1}^{J} D(\{P_j[k|x]\}||\{P[k|x]\}) \tag{1}$$

iii) Compare (as best you can) the solutions in parts i) and ii). Can you imagine a situation where one of the estimators may be preferred ?