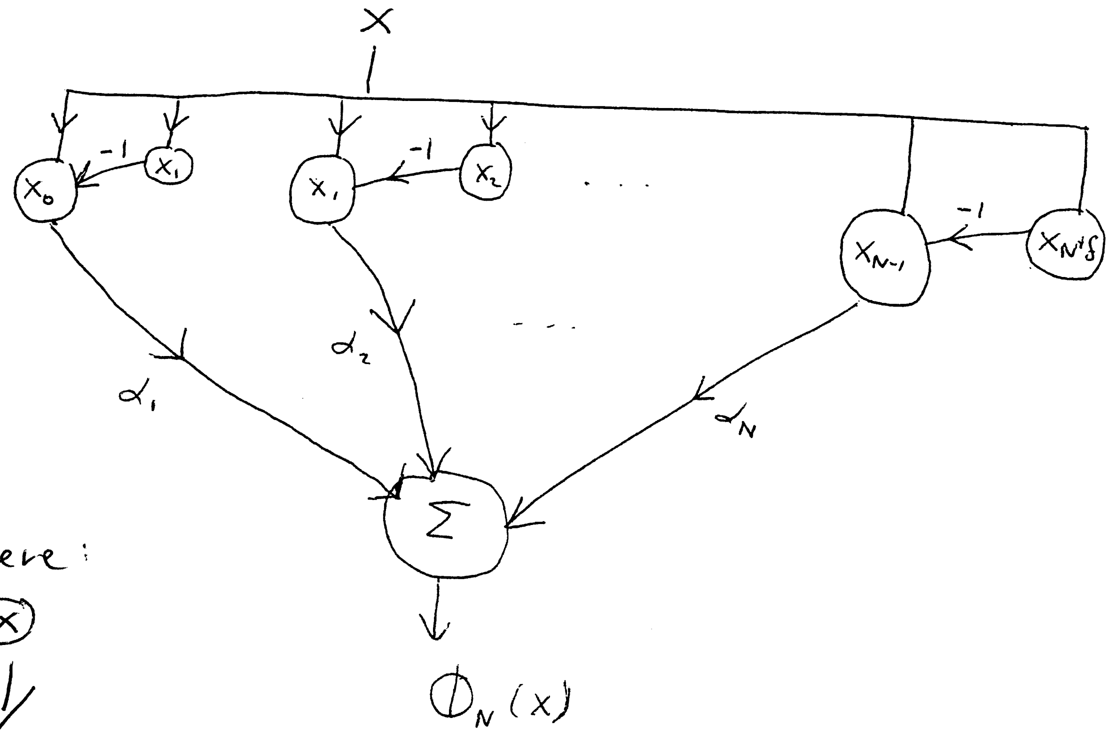
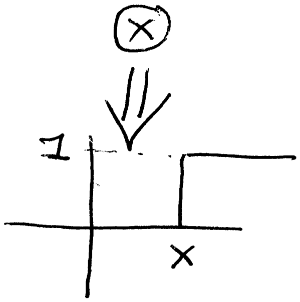
HW #5 Solutions,
EE 556,

#1)



where:



#2) $f_{com}(\underline{x}) = \sum_{i=1}^{M} w_i f_i(\underline{x})$, where $\sum_{i=1}^{M} w_i = 1$

Recall from lecture, for the standard committee machine, that we wrote the MSE in the form:

$$\frac{1}{T}\cdot\frac{1}{N^2}(1\,1\,1\cdots 1)EE^T(1\,1\,1\cdots 1)^T$$

For the weighted committee machine structure, easy to show that $MSE = \frac{1}{T}\cdot\frac{1}{N^2}\underline{w}EE^T\underline{w}^T$,

where $\underline{w} = (w_1\ w_2\ \cdots\ w_M)$

Thus, the Lagrangian cost function is:

$$L = \frac{1}{T}\cdot\frac{1}{N^2}\underline{w}EE^T\underline{w}^T + \lambda(\underbrace{(1\,1\cdots 1)}_{\underline{1}}\underline{w}^T - 1)$$

$$\nabla_{\underline{w}} L = \frac{2}{T}\cdot\frac{1}{N^2}\underline{w}EE^T + \lambda\underline{1} = 0$$

<u>Assuming EE^T is invertible</u>, this gives

$$\underline{w} = -\lambda\frac{N^2}{2}T\underline{1}(EE^T)^{-1}$$

Also, from the constraint:

$$\underline{w}\,\underline{1}^T = -\lambda\frac{N^2}{2}T\underline{1}(EE^T)^{-1}\underline{1}^T = 1 \Rightarrow$$

$$\lambda = \frac{-1}{\frac{N^2}{2}T\underline{1}(EE^T)^{-1}\underline{1}^T}, \quad \text{i.e.,}$$

$$\underline{w} = \frac{\underline{1}(EE^T)^{-1}}{\underline{1}(EE^T)^{-1}\underline{1}^T}$$

This problem is not so well-stated.

(5.12, Haykin) Let's assume that the training set consists of $\{(\underline{x}_i, f(\underline{x}_i)), i=1,\dots N\}$.

However, <u>in practice</u>, the regression function we design will not have as input $\underline{x}$, but rather $\underline{x} + \underline{\varepsilon}$, where $\underline{\varepsilon}$ is additive noise. We'd like to take into account the additive noise <u>in an optimal way</u>, in designing the regression function.

To do so, we need to minimize the <u>expected squared</u> error:

$$J(F) = \frac{1}{2}\sum_{i=1}^{N} \int_{\mathbb{R}^{m_0}} \left(f(\underline{x}_i) - F(\underline{x}_i + \underline{\varepsilon})\right)^2 f_{\underline{\varepsilon}}(\underline{\varepsilon})\, d\underline{\varepsilon}$$

Let $\underline{z} = \underline{x}_i + \underline{\varepsilon}$
$d\underline{z} = d\underline{\varepsilon}$

$$= \frac{1}{2}\sum_{i=1}^{N} \int_{\mathbb{R}^{m_0}} \left(f(\underline{x}_i) - F(\underline{z})\right)^2 f_{\underline{\varepsilon}}(\underline{z} - \underline{x}_i)\, d\underline{z}$$

$$= \int_{\mathbb{R}^{m_0}} \frac{1}{2}\sum_{i=1}^{N} \left(f(\underline{x}_i) - F(\underline{z})\right)^2 f_{\underline{\varepsilon}}(\underline{z} - \underline{x}_i)\, d\underline{z}$$

Now, let $f_M(\underline{z}) = \dfrac{\sum_{i=1}^{N} f(\underline{x}_i) f_{\underline{\varepsilon}}(\underline{z} - \underline{x}_i)}{\sum_{i=1}^{N} f_{\underline{\varepsilon}}(\underline{z} - \underline{x}_i)}$ and apply our "usual" trick

Then, $$J(F) = \int_{\mathbb{R}^{m_0}} \frac{1}{2}\sum_{i=1}^{N} \left(\left(f(\underline{x}_i) - f_M(\underline{z})\right) + \left(f_M(\underline{z}) - F(\underline{z})\right)\right)^2 f_{\underline{\varepsilon}}(\underline{z} - \underline{x}_i)\, d\underline{z}$$

$$= \int_{\mathbb{R}^{m_0}} \frac{1}{2}\sum_{i=1}^{N} \left(f(\underline{x}_i) - f_M(\underline{z})\right)^2 f_{\underline{\varepsilon}}(\underline{z} - \underline{x}_i)\, d\underline{z} + \int_{\mathbb{R}^{m_0}} \frac{1}{2}\sum_{i=1}^{N} \left(f_M(\underline{z}) - F(\underline{z})\right)^2 \cdot f_{\underline{\varepsilon}}(\underline{z} - \underline{x}_i)\, d\underline{z}$$

The second term is minimized by choosing $F(\underline{z}) = f_M(\underline{z})$.

#4) Consider an RBF discriminant function:

$$g_K(\underline{x}) = \sum_{l=1}^{M} \lambda_{lk}\left(\frac{f(\underline{x};\theta_l)}{\sum_{m=1}^{M} f(\underline{x};\theta_m)}\right), \quad \text{where}$$

we assume $f(\cdot;\theta)$ is a density function.

The decision rule is:

$$K^* = \arg\max_K g_K(\underline{x}) = \arg\max_K \sum_{l=1}^{M} \lambda_{lk} f(\underline{x};\theta_l)$$

We want to show that this rule is equivalent to the Bayes decision rule associated with a particular statistical mixture model.

To show this, first define $\lambda_{min} = \min_{(l,k)} \lambda_{lk}$

Then, subtracting $\lambda_{min} \sum_{k'} f(\underline{x};\theta_{k'})$ from $\sum_{l=1}^{M} \lambda_{lk} f(\underline{x};\theta_l)$ we get the equivalent rule:

$$K^* = \arg\max_K \sum_l f(\underline{x};\theta_l)\tilde{\lambda}_{lk}, \quad \text{where } \tilde{\lambda}_{lk} = \lambda_{lk} - \lambda_{min}.$$

Note that $\tilde{\lambda}_{lk} \geq 0$. Next, divide each discriminant fn. by $\left(\sum_{(m,n)} \tilde{\lambda}_{mn}\right)$, to get

$$K^* = \arg\max_K \sum_l f(\underline{x};\theta_l) q_{lk}, \quad \text{where } q_{lk} = \frac{\tilde{\lambda}_{lk}}{\sum_{mn} \tilde{\lambda}_{mn}}$$

Note: $0 \leq q_{lk} \leq 1$ + $\sum_{lk} q_{lk} = 1.$

Finally, divide each discriminant function by the sum: $\sum_{k',l} f(\underline{x};\theta_l)\, q_{lk'}$ to yield the equivalent rule:

$$k^* = \arg\max_k \frac{\sum_l f(\underline{x};\theta_l)\, q_{lk}}{\sum_{k',l} f(\underline{x};\theta_l)\, q_{lk'}}$$

$$= \arg\max_k \frac{\sum_l f(\underline{x};\theta_l) \frac{q_{lk}}{\sum_{k'} q_{lk'}} \cdot \left(\sum_{k'} q_{lk'}\right)}{\sum_l f(\underline{x};\theta_l)\left(\sum_{k'} q_{lk'}\right)} \quad (*)$$

Note that since $\sum_{k',l} q_{lk'} = 1 \implies$

$$\sum_l \left(\sum_{k'} q_{lk'}\right) = 1$$

This gives $\sum_{k'} q_{lk'}$ the interpretation of a "component mass".

Now, consider the following statistical mixture model for generating the pair $(\underline{x}, c)$:

1) Randomly select one of M components, based on component masses $\{\alpha_k, k=1..M\}$.
2) Given the selected component j, randomly select $\underline{x}$ according to $f(\underline{x};\theta_j)$.
3) Randomly select the class c according to the component conditional probability mass function $P[C=m / J=j]$.

The associated class posterior is:

$$\mathrm{Prob}[C=c/\underline{X}=\underline{x}] = \frac{\sum_{l} \alpha_l f(\underline{x};\Theta_l)\,\mathrm{Prob}[C=c/J=l]}{\sum_{l} \alpha_l f(\underline{x};\Theta_l)} \quad (\Delta)$$

Note that the posterior in $(*)$ has the same form as the posterior in (Δ).

Thus, an RBF classifier is equivalent to a Bayes classifier for a special statistical mixture model.

This result was shown in (Miller-Uyar, Neural Computation, 1998).

#5) Let's view the RBF from the viewpoint of a stochastic model.

Imagine the feature vector $\underline{X} = \underline{x}$ has already been generated. Consider randomly selecting a basis function, given $\underline{X} = \underline{x}$, i.e. according to the posterior probability

$$\text{Prob}[J = j / \underline{X} = \underline{x}] = \frac{f(\underline{x}; \theta_j)}{\sum_{l=1}^{m} f(\underline{x}; \theta_l)}$$

and then <u>deterministically</u> producing $Y = \lambda_j$

In this case, the RBF output can be interpreted as a <u>conditional mean estimator</u>,

i.e.

$$E[Y / \underline{X} = \underline{x}] = \sum_{j=1}^{m} \text{Prob}[J = j / \underline{X} = \underline{x}] \cdot E[Y / J = j]$$

But since $f_{Y/j}(y) = \delta(y - \lambda_j)$, $E[Y/J=j] = \lambda_j$,

i.e.

$$E[Y / \underline{X} = \underline{x}] = \sum_{j=1}^{m} \left(\frac{f(\underline{x}; \theta_j)}{\sum_{l=1}^{m} f(\underline{x}; \theta_l)} \right) \cdot \lambda_j$$

<u>Note</u>: if the statistical model is accurate, the conditional mean estimator is the optimal (minimum mean-squared error) estimator.

#6) Suppose we define the Gaussian basis function as $f(\underline{x};\theta_j) = \frac{1}{\sqrt{2\pi}^d \sigma^{d/2}} \cdot \exp\left(-\frac{\|\underline{x} - \underline{c}_j\|^2}{2\sigma^2}\right)$ instead of just $\exp\left(-\frac{\|\underline{x} - \underline{c}_\ell\|^2}{2\sigma^2}\right)$.

The RBF becomes:

$$f_{RBF}(\underline{x}) = \sum_{\ell=1}^{M} \lambda_\ell \cdot \frac{1}{\sqrt{2\pi}^d \sigma^{d/2}} \cdot \exp\left(\frac{-\|\underline{x} - \underline{c}_\ell\|^2}{2\sigma^2}\right)$$

$$= \sum_{\ell=1}^{M} \tilde{\lambda}_\ell \exp\left(\frac{-\|\underline{x} - \underline{c}_\ell\|^2}{2\sigma^2}\right)$$

In other words, since anyway we optimize over the weights $\{\lambda_\ell\}$ as free parameters, use of the width parameter σ in defining the basis function does not add any representation power.

This statement in fact also holds if the widths are basis-function-dependent, i.e. σ_ℓ.

#7)

i) $$L = \frac{1}{J}\sum_{j=1}^{J}\sum_{k} P[k/x] \log \frac{P[k/x]}{P_j[k/x]} + \lambda\left(\sum_{k} P[k/x] - 1\right)$$

$$\frac{\partial L}{\partial P[m/x]} = \frac{1}{J}\sum_{j=1}^{J}\left(\log \frac{P[m/x]}{P_j[m/x]} + 1\right) + \lambda = 0 \Rightarrow$$

$$\log P[m/x] = \frac{1}{J}\sum_{j=1}^{J} \log P_j[m/x] - 1 - \lambda$$

$$\Rightarrow P[m/x] = \left(\prod_{j=1}^{J} P_j[m/x]\right)^{1/J} e^{-1-\lambda}$$

Imposing the pmf constraint (through the choice of λ) we have:

$$P[m/x] = \frac{\left(\prod_{j=1}^{J} P_j[m/x]\right)^{1/J}}{\sum_{k}\left(\prod_{j=1}^{J} P_j[k/x]\right)^{1/J}}$$

This is a "geometric averaging" rule.

ii) $$L = \frac{1}{J}\sum_{j=1}^{J}\sum_{k} P_j[k/x] \log \frac{P_j[k/x]}{P[k/x]} + \lambda\left(\sum_{k} P[k/x] - 1\right)$$

$$\frac{\partial L}{\partial P[m/x]} = -\frac{1}{J}\sum_{j=1}^{J} \frac{P_j[m/x]}{P[m/x]} + \lambda = 0 \Rightarrow$$

$$P[m/x] = \frac{1}{J}\sum_{j=1}^{J} P_j[m/x]$$

Note that $P[m/x]$ is a valid probability mass function.

iii) geometric average may be more sensitive to an unreliable expert, e.g. consider $P_j[C = c_{true}/x] = 0$ -- in this case the ensemble decision ~~must~~ must be incorrect...