CS 542, Spring 2014

Assignment 3

Shan Sikdar

1 4.2

Show that as a consequence of this contraint, the elements of the model prediction y(x) given by the least-squares solution (4.17) also satisfy this constraint, so that $a^T y(x) + b = 0$ To do so assume that one of the basis functions $\phi_0(x) = 1$ so that the corresponding w_0 plays the role of a bias.

The sum-of-square-errors function is very similar to the one that was used in chapter 3. So similar to how the bias parameter was made explicit in (3.18), the bias term in the sum-of-squares error function can be made explicit as well.
So rewrite:

$$E_D(W^{\sim}) = \frac{1}{2} Tr\{(X^{\sim}W^{\sim} - T)^T (X^{\sim}W^{\sim} - T)\}$$

as this: $E_D(W^{\sim}) = \frac{1}{2} Tr\{(XW + 1w_0^T - T)^T (XW + 1w_0^T - T)\}$
Use $\frac{d}{dA} Tr(A^T A) = 2A^T$.
So we now hace the fomrula: $2 * (XW + 1w_0^T - T)^T (1) = 2Nw_0 + 2(XW - T)^T 1 = 0$
So $w_0 = \frac{1T^T}{N} 1 - \frac{1(XW)^T}{N}$
$w_0 = \frac{1T^T}{N} 1 - \frac{(W^T X^T)1}{N}$
If we defined a $\bar{t} = \frac{1T^T}{N} 1$ and $\bar{x} = \frac{X^T 1}{N}$similar to the style of equation (3.20).
We can rewrite: $w_0 = \bar{t} - W^T \bar{x}$
Now that w_0 has been properly defined, it can be plugged back into the

sum of squares error function.
$E_D(W^{\sim}) = \frac{1}{2}Tr\{(XW + 1w_0^T - T)^T(XW + 1w_0^T - T)\}$
$= E_D(W^{\sim}) = \frac{1}{2}Tr\{(XW+1(\bar{t}-W^T\bar{x})^T-T)^T(XW+1(\bar{t}-W^T\bar{x})^T-T)\}$
Let $\bar{T} = 1\bar{t}$ and $\bar{X} = 1\bar{x}$:
$= E_D(W^{\sim}) = \frac{1}{2}Tr\{(XW + \bar{T} - \bar{X}W - T)^T(XW + \bar{T} - \bar{X}W - T)\}$

Taking the derivative with respect to W. Use $\frac{d}{dA}Tr(A^TA) = 2A^T$.
So we get:
$\nabla_W = 0 = 2(XW - T)X^T$
Solving for W similar to how we did in lab. We get:
$W - (X^TX)^{-1}X^TT.$
These wieghts make up our new a for the equation (4.157). And our w_0 is our b for equation (4.157) if we take an arbitary vector produced by this function, tranform it by y(x) and plug into into equation y(x) we get the 0 vector. Therefore the contstraint is satisfied.

2 4.10

The likelihood function for k classes can be derived by extending the 2 case class example we did in class:
$P(\{\phi_n, t_n|\{\pi_k\}\}) = \Pi_{n=1}^{N}\Pi_{k=1}^{K}\{p(\phi_n|C_k)\pi_k\}^{t_{nk}}$
Log of the likelihood:
$ln(P(\{\phi_n, t_n|\{\pi_k\}\})) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\{lnp(\phi_n|C_k) + ln\pi_k\}$
Given assumption for problem 4.10: $p(\phi|C_k) = N(\phi|\mu, \Sigma)$ So the log of likelihood becomes:
$ln(P(\{\phi_n, t_n|\{\pi_k\}\})) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\{lnN(\phi|\mu, \Sigma) + ln\pi_k\}$
Using the format from lab, set $\alpha = \frac{1}{(2\pi)^{.5}}$ and $\beta = \frac{1}{|\Sigma|^{.5}}$:
$ln(P(\{\phi_n, t_n|\{\pi_k\}\})) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\{ln\alpha+ln\beta+(\frac{-1}{2}(\phi_n-\mu_k)^T\Sigma^{-1}(\phi_n-\mu_k)) + ln\pi_k\}$
Note that α and π_k is just a constant so it will always drop out no matter what analysis we do:
$ln(P(\{\phi_n, t_n|\{\pi_k\}\})) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\{ln\beta+(\frac{-1}{2}(\phi_n-\mu_k)^T\Sigma^{-1}(\phi_n-\mu_k))\}$

Taking the derivative with respect to μ_k. Using an identity found where $\frac{d}{ds}(x-s)^T W(x-s) = -2W(x-s)$.
We get:
$0 = \nabla_{\mu_k} = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\Sigma^{-1}(\phi_n - \mu_k)$
$0 = \nabla_{\mu_k} = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\phi_n - \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\mu_k$
$\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\phi_n = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\mu_k$
Using the fact that if I sum over n all $t_k n$ I will get the number of instances for a paticular class aka N_k:
$\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\phi_n = \sum_{k=1}^{K} N_k\mu_k$
$\sum_{n=1}^{K}\sum_{k=1}^{N} \frac{1}{N_k} t_{nk}\phi_n = \sum_{k=1}^{K} \mu_k$
$\sum_{k=1}^{N} \frac{1}{N_k} t_{nk}\phi_n = \mu_k$

Taking the derivative with respect to the shared covariance:
Look at all the parts of the log of the likelihood that depend on Σ:

$ln(P(\{\phi_n, t_n|\{\pi_k\}\})) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\{ln\beta+(\frac{-1}{2}(\phi_n-\mu_k)^T\Sigma^{-1}(\phi_n-\mu_k))\}$

$= \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\{\frac{-1}{2}ln|\Sigma| + (\frac{-1}{2}(\phi_n - \mu_k)^T\Sigma^{-1}(\phi_n - \mu_k))\}$
$= \frac{-1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\{ln|\Sigma| + ((\phi_n - \mu_k)^T\Sigma^{-1}(\phi_n - \mu_k))\}$
$= \frac{-1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}ln|\Sigma| + \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}((\phi_n - \mu_k)^T\Sigma^{-1}(\phi_n - \mu_k))\}$
$= \frac{-1}{2}\sum_{k=1}^{K} N_k ln|\Sigma| + \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}((\phi_n - \mu_k)^T\Sigma^{-1}(\phi_n - \mu_k))$
$= \frac{-N}{2}ln|\Sigma| + \frac{-1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}Tr((\Sigma^{-1}(\phi_n - \mu_k)^T(\phi_n - \mu_k)))$
$= \frac{-N}{2}ln|\Sigma| + \frac{-1}{2}Tr((\Sigma^{-1}\sum_{k=1}^{K}\sum_{n=1}^{N} t_{nk}(\phi_n - \mu_k)^T(\phi_n - \mu_k)))$
$= \frac{-N}{2}ln|\Sigma| + \frac{-1}{2}Tr(\Sigma^{-1}S)$, where S is defined similar to lab except instead of having two classes we have K classes:
$S = \sum_{k=1}^{K}\sum_{n=1}^{N} t_{nk}(\phi_n - \mu_k)^T(\phi_n - \mu_k)$
So now taking derivative with respect to Σ. (Using the same method used in the lab session)
$0 = \nabla_{\mu_k} = \frac{-N}{2}(\Sigma^{-1})^T - \frac{1}{2}(\Sigma^{-1}S\Sigma^{-1})^T$
$\Sigma = \frac{1}{N}S$
$\Sigma = \frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{N} t_{nk}(\phi_n - \mu_k)^T(\phi_n - \mu_k)$
$\Sigma = \sum_{k=1}^{K}\frac{N_k}{N}\frac{1}{N_k}\sum_{n=1}^{N} t_{nk}(\phi_n - \mu_k)^T(\phi_n - \mu_k)$
Which is the result if you combine 4.162 and 4.163 on page 222.

3 4.14

Since the data is linearly seperable, the boundary line will be when $w^T\phi(x) = 0$. Minimizing the error of 4.90 is the same as maximizing the likelihood answer. This will happen when the derivative (equation 4.91) is 0. But this will only happen when $y_n = t_n$. But y_n is on the descision boundary so its zero. But $y_n = \sigma(w^T\phi(x)) = 0$ But in order for $\sigma(w^T\phi(x))$ to be 0 w^T must increase infinietely (since ϕ is fixed) and therefore the magnitude of w is taken to infinity.