CS 542, Spring 2014

Assignment 2

Shan Sikdar

1 3.3

Find an expression for the solution w* that minimizes this error function. Give two alternative interpretations of the weighted sum-of-errors function in terms of (i) data dependant noise and (ii) replicated data points.

Take the new sum -of-squares error function:
$E_D(w) = \frac{1}{2}\sum_{n=1}^{N} r_n\{t_n - w^T\phi(x_n)\}^2$
$\nabla E_D(w) = \sum_{n=1}^{N} r_n\{t_n - w^T\phi(x_n)\}\phi(x_n)^T$
$0 = \sum_{n=1}^{N} r_n t_n \phi(x_n)^T - \sum_{n=1}^{N} w^T\phi(x_n)\ \phi(x_n)^T$
$\sum_{n=1}^{N} w^T\phi(x_n)\ \phi(x_n)^T = \sum_{n=1}^{N} r_n t_n \phi(x_n)^T$
Since $\phi = (\phi_0, \phi_1...., \phi_{m-1})^T$:
$\sum_{n=1}^{N} w^T(\phi_0, \phi_1...., \phi_{m-1})((\phi_0, \phi_1 m...., \phi_{m-1})^T)^T = \sum_{n=1}^{N} r_n t_n \phi(x_n)^T$
Now $r_1, ..., r_n$ can be represented by a diagnol matrix where the coefficents make up the diaganol and everywhere else is 0.
Then using that fact along with the fact that summing $((\phi_0, \phi_1 m...., \phi_{m-1})^T)^T$ from 1 to N gives you Φ you get:
$w^T\Phi^T R\Phi = \sum_{n=1}^{N} r_n t_n \phi(x_n)^T$
$w^T\Phi^T R\Phi = \sum_{n=1}^{N} r_n t_n((\phi_0, \phi_1 m...., \phi_{m-1})^T)^T$
$w^T = \Phi^T Rt$
$w^T = \Phi^{-1}R^{-1}(\Phi^T)^{-1}\Phi^T Rt$
$w^T = (R\Phi)^{-1}(\Phi^T)^{-1}\Phi^T Rt$
$w^* = (\Phi^T R\Phi)^{-1}\Phi^T Rt$
(i) In terms of data dependent noise variation r_n can be seen as an inverse varianve parameter to a datapoint (x_n, t_n) which modifies the precision matrix.

(ii) replicated data points: r_n can be regarded as an effective number of replicated observations of a data point (x_n, t_n).

2 3.11

Show that the uncertainty $\sigma_N(x)$ assocaited with linear regression function given by (3.59) statisfies $\sigma_{N+1}(x) \leq \sigma_N(x)$
Fromula 3.59: $\sigma_N^2(x) = \frac{1}{\beta} + \phi(x)^T S_N \phi(x)$
The formula for 3.59 for N+1 case:
$\sigma_{N+1}^2(x) = \frac{1}{\beta} + \phi(x)^T S_{N+1} \phi(x)$
It turns out from problem 3.8 the posterior distribution is given by 3.49 but with S_N replaced with S_{N+1} and m_n replaced with m_{n+1}. So:
$S_{N+1}^{-1} = S_N^{-1} + \beta\phi_{N+1}\phi_{N+1}^T$
$S_{N+1} = (S_N^{-1} + \beta\phi_{N+1}\phi_{N+1}^T)^{-1}$
Then using the matrix identity from appendix C: where $M = S_N^{-1}$ and $v = \beta^{\frac{1}{2}}\phi_{N+1}$:
$S_{N+1} = S_N - \frac{(S_N\phi_{N+1}\beta^{\frac{1}{2}})(\beta^{\frac{1}{2}}\phi_{N+1}^T S_N)}{1+\beta\phi_{N+1}S_N\phi_{N+1}}$
$S_{N+1} = S_N - \frac{\beta(S_N\phi_{N+1})(\phi_{N+1}^T S_N)}{1+\beta\phi_{N+1}S_N\phi_{N+1}}$
So plugging this value of S_{N+1} back into formula 3.59:
$\sigma_{N+1}^2(x) = \frac{1}{\beta} + \phi(x)^T(S_N - \frac{\beta(S_N\phi_{N+1})(\phi_{N+1}^T S_N)}{1+\beta\phi_{N+1}S_N\phi_{N+1}})\phi(x)$
$\sigma_{N+1}^2(x) = \frac{1}{\beta} + \phi(x)^T S_N \phi(x) - \frac{\phi(x)^T\beta(S_N\phi_{N+1})(\phi_{N+1}^T S_N)\phi(x)}{1+\beta\phi_{N+1}S_N\phi_{N+1}}$
Using the definition of σ_N^2 from above:
$\sigma_{N+1}^2(x) = \sigma_N^2 - \frac{\phi(x)^T\beta(S_N\phi_{N+1})(\phi_{N+1}^T S_N)\phi(x)}{1+\beta\phi_{N+1}S_N\phi_{N+1}}$
Now if the fraction on the left is always positive then we are done. Since S_N is postive definite, the numerator and demoniator will both be positive.
Therefore:
$\sigma_{N+1}(x) \leq \sigma_N(x)$

3 3.14

Show for $\alpha = 0$, the equivelent kernel cna be writtn as $k(x, x') = \psi(x)^T \psi(x')$
Using equation 3.54, when $\alpha = 0$, $S_N^{-1} = \beta \Phi^T \Phi$
Since $\psi_j(x)$ is our new othronormal basis passning the same space, there must be some function that can transform the original basis, to the orthonormal one. So in other words:
$\psi(x) = V\phi(x)$
where V is the matrix that represents the function that transforms our original basis to the new one. (similar to gram-schmidt?). Since the result of the transformation cover the same space, there must be another function that takes the orthonormal one back to the original space. Therefore this matrix must be the inverse of V. Then V has an inverse.
Before I can work with wquation 3.54, I first need to transform Φ into the new orthonormal basis so:
$\Phi V^T = \Psi$ (using transpose since multiplying on left instead of right of Φ)
Solving back for Φ, $\Phi = \Psi(V^T)^{-1}$:
So going back to equation 3.54:
$S_N^{-1} = \beta \Phi^T \Phi$
$S_N = (\beta \Phi^T \Phi)^{-1}$
$S_N = \beta^{-1}(\Phi^T \Phi)^{-1}$
$= \beta^{-1}(((\Psi(V^T)^{-1})^T(\Psi(V^T)^{-1}))^{-1}$
$= \beta^{-1}(\Psi(V^T)^{-1})^{-1}(((\Psi(V^T)^{-1})^T)^{-1}$
$= \beta^{-1}((V^T)^{-1})^{-1}\Psi^{-1})(((V^T)^{-1})^T\Psi^T)^{-1}$
$= \beta^{-1}((V^T))\Psi^{-1})(((V^T)^{-1})^T\Psi^T)^{-1}$
$= \beta^{-1}((V^T))\Psi^{-1})(((\Psi^T)^{-1}((V^T)^{-1})^T)^{-1}$
$= \beta^{-1}((V^T))\Psi^{-1})(\Psi^T)^{-1}V) = \beta^{-1}((V^T))V)$
(Ψ is an ortho basis, transpose and inverse will negate each other.)
$= \beta^{-1}((V^T))V)$
Plugging in S_N into the equivelent kernel formula:
$K(x, x') = \beta\phi(x)^T S_N \phi(x') = \beta\phi(x)^T \beta^{-1}((V^T))V)\phi(x') = \phi(x)^T((V^T))V)\phi(x')$
Now applying ϕ to the linear transformation V and V^T:
$= \psi(x)^T\psi(x')$
Now looking at the summation of the equivelence kernels:
$\sum_{n=1}^N k(x, x_n) = \sum_{n=1}^N \psi(x)^T\psi(x_n) = \sum_{n=1}^N \sum_{i=1}^M \psi_i(x)^T\psi_i(x_n)$
Using 3.115: $= \sum_{i=1}^M \psi_i(x_n)\psi_i(x_n) = 1$

4 3.21

Prove 3.117 and then make use of 3.117 to derive 3.92 starting drom 3.86.
Let A be a real, symetric matrix A.
By definition eigen values and vectors define how a vetor under the transformation of A can be recreated by simply multiplying the original vector by some constant λ_i(the eigenvalues). Let $\{u_i\}$ be a set of orthonormal vectors then we know be defintion of eignvalues/vectors:
$Au_i = \lambda_i u_i$. Then we can also define A by its eigenvalues.
So:
$ln|A| = ln\Pi_{i=1}^{M}\lambda_i = \sum_{i=1}^{M} ln\lambda_i$
Let α be some random optimized value:

$\frac{\delta ln|A|}{\delta\alpha} = \sum_{i=1}^{M} \frac{1}{\lambda_i}\frac{\delta\lambda_i}{\delta\alpha}$
Now show that the right hand side of equation 3.117 is equal to this.
From C.45 and C.46 in the Appendix on matrices, given eigenvalues we can recreate matrices and their inverse with their eigenvalues:
$A = \sum_{i=1}^{M} \lambda_i u_i u_i^T$ (c.45)
$A^{-1} = \sum_{i=1}^{M} \frac{1}{\lambda_i} u_i u_i^T$(c.46)
Taking the derivative of A using the product rule:
$\frac{\delta}{\delta\alpha}A = \sum_{i=1}^{M} \frac{\delta\lambda_i}{\delta\alpha} u_i u_i^T + \lambda_i(\frac{\delta u_i}{\delta\alpha} u_i^T + u_i \frac{\delta u_i}{\delta\alpha}^T)$
If the length of u_i is always constant this implies the derivative of a vector is orthognal to the vector . Then the multiplication u_i and $(\frac{\delta u_i}{\delta\alpha})$ will be zero. ("Courtesy of http://www.physicsforums.com/showthread.php?t=523876" and some calc III). So the last term cancels out leaving:
$\frac{\delta}{\delta\alpha}A = \sum_{i=1}^{M} \frac{\delta\lambda_i}{\delta\alpha} u_i u_i^T$
Now plugging in A^{-1}, $\frac{\delta}{\delta\alpha}A$ ino the right hand side of 3.117:
$Tr(A^{-1}\frac{\delta}{\delta\alpha}A) = Tr(\sum_{i=1}^{M} \frac{1}{\lambda_i} u_i u_i^T \sum_{j=1}^{M} \frac{\delta\lambda_j}{\delta\alpha} u_j u_j^T)$
$= tr(\sum_{i=1}^{M}\sum_{j=1}^{M} \frac{1}{\lambda_i}\frac{\delta\lambda_j}{\delta\alpha} u_i u_i^T u_j u_j^T)$
Now using the fact that the multipcation of orthognal vectors will be 0 leaves: $Tr(\frac{1}{\lambda_i}\frac{\delta\lambda_j}{\delta\alpha}\sum_{i=1}^{M} u_i u_i^T)$
But $\sum_{i=1}^{M} u_i u_i^T = I$ So we we are left with: $Tr((\frac{1}{\lambda_i}\frac{\delta\lambda_j}{\delta\alpha}) = \sum_{i=1}^{M}\frac{1}{\lambda_i}\frac{\delta\lambda_i}{\delta\alpha}$
Since Both sides of equation 3.117 come out to $\sum_{i=1}^{M}\frac{1}{\lambda_i}\frac{\delta\lambda_i}{\delta\alpha}$ the must be equal and therefore the equation holds.
To prove 3.92 from starting with 3.86:
$ln(t|\alpha,\beta) = \frac{M}{2}ln\alpha + \frac{N}{2}ln\beta - E(m_N) - \frac{1}{2}ln|A| - \frac{N}{2}ln(2\pi)$

$\frac{dln(t|\alpha,\beta)}{\alpha} = \frac{M}{2}\frac{1}{\alpha} - \frac{1}{2}m_N^T m_N - \frac{1}{2}Tr(A^{-1}\frac{d}{d\alpha}A)$
$= \frac{1}{2}(\frac{M}{\alpha} - m_N^T m_N - Tr(A^{-1}\frac{d}{d\alpha}A))$
Since $A = S_N^{-1}$ it represents the mean of the prior distribution. In otherwords it is α So $\frac{d}{d\alpha}A = 1$
$= \frac{1}{2}(\frac{M}{\alpha} - m_N^T m_N - Tr(A^{-1}))$
$= \frac{1}{2}(\frac{M}{\alpha} - m_N^T m_N - \sum \frac{1}{\lambda_i+\alpha}))$
This is the right hand side of equation 3.89. Since 3.89 is used to derive 3.92, then 3.117 can be used to derive 3.92 using 3.89 as an intermdiary step.