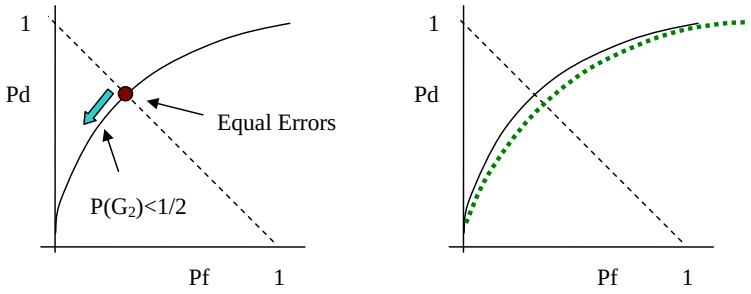
EE 511 – Homework 3 Solutions

1. The ROC curve plots P_d vs. P_f as illustrated in the left figure below. The equal error point is where $P_m = 1 - P_d = P_f$; the dashed line illustrated in the figure shows this function and the dot at its intersection with the ROC curve is the equal error point.

Figure 1: Rough ROC curve sketch.



When $P(G_1) > P(G_2)$ then you tend to decide G_1 more often in a minimum error rate decision (i.e. the G_1 decision region is bigger), so $P_d = \int_{D_2} p(x|G_2)dx$ is smaller, as is P_f. Therefore, the minimum error rate point would shift to the left (lower P_d, lower P_f) as indicated by the arrow. Note that changing the priors does not change the ROC curve. It DOES change the location on the ROC curve where the total error $P_f P(G_1) + P_m P(G_2)$ is minimum, so it would change the decision boundary if you were selecting it based on that criterion, and would therefore change the values of P_f and P_m for your decision boundary.

If we observed $X + N$ instead, the variance would be higher and for a given P_f we would have a lower P_d. The ROC curve would flatten out more, as in the dotted curve in the right figure.

2. The MAP rule select the class according to:

$$i^* = \underset{i}{\operatorname{argmax}}\, P(i)p(x|i) = \underset{i}{\operatorname{argmax}}\, p(x|i)$$

since all $P(i)$ are equal. But

$$p(x|i) = \frac{1}{(2\pi)^{d/2}\sigma^d} e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{d}(x_j - \mu_{ij})^2},$$

where x_j and μ_{ij} are the jth components of x and μ_i. Therefore we want:

$$\begin{aligned} i^* &= \underset{i}{\text{argmax}} \log p(x|i) \\ &= \underset{i}{\text{argmax}} \left[K - \frac{1}{2\sigma^2} \sum_{j=1}^{d} (x_j - \mu_{ij})^2 \right] \\ &= \underset{i}{\text{argmin}} \sum_{j=1}^{d} (x_j - \mu_{ij})^2 \end{aligned}$$

which is the minimum Euclidean distance to the mean.

3. HTF 3.6: Show that the ridge regression estimate is the mean (and mode) of the posterior distribution under Gaussian...

   The likelihood: $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2 I)$. The prior: $\beta \sim N(0, \tau^2 I)$. Assume that our input data x is already centered (see HTF pg. 59) so that we can ignore the intercept term β_0. The posterior distribution is:

$$p(\beta|\mathbf{y}, \mathbf{X}) = \frac{1}{Z} p(\mathbf{y}|\beta, \mathbf{X}) p(\beta)$$

   where $Z = Z(\mathbf{y}, \mathbf{X}) = \int p(\mathbf{y}|\beta, \mathbf{X}) p(\beta) d\beta$ is the normalization constant that does not dependent on β.

   Writing out the (log) posterior distribution explicitly:

$$\begin{aligned} p(\beta|\mathbf{y}, \mathbf{X}) &= \frac{1}{Z} p(\mathbf{y}|\beta, \mathbf{X}) p(\beta) \\ &= \frac{1}{Z} \frac{1}{(2\pi)^{p/2}\sigma^p} \exp\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^t(\mathbf{y} - \mathbf{X}\beta)\} \frac{1}{(2\pi)^{p/2}\tau} \exp\{-\frac{\beta^t\beta}{2\tau^2}\} \\ \log p(\beta|\mathbf{y}, \mathbf{X}) &= -\log Z - K - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^t(\mathbf{y} - \mathbf{X}\beta) - \frac{\beta^t\beta}{2\tau^2} \end{aligned}$$

   where K corresponds to the other terms outside the exp. Taking the derivative of the $\log p(\beta|\mathbf{y}, \mathbf{X})$ wrt β, we have:

$$\frac{d \log p(\beta|\mathbf{y}, \mathbf{X})}{d\beta} = \frac{1}{\sigma^2} \mathbf{X}^t(\mathbf{y} - \mathbf{X}\beta) - \frac{\beta}{\tau^2} \tag{1}$$

Setting this to zero,

$$\begin{aligned} 0 = & \frac{1}{\sigma^2}\mathbf{X}^t(\mathbf{y} - \mathbf{X}\beta) - \frac{\beta}{\tau^2} \\ \rightarrow \quad 0 = & \mathbf{X}^t(\mathbf{y} - \mathbf{X}\beta) - \frac{\sigma^2}{\tau^2}\beta \\ \rightarrow \quad 0 = & \mathbf{X}^t\mathbf{y} - \mathbf{X}^t\mathbf{X}\beta - \frac{\sigma^2}{\tau^2}\beta \\ \rightarrow \quad 0 = & \mathbf{X}^t\mathbf{y} - (\mathbf{X}^t\mathbf{X} + \frac{\sigma^2}{\tau^2}I)\beta \\ \rightarrow \quad \hat{\beta} = & (\mathbf{X}^t\mathbf{X} + \frac{\sigma^2}{\tau^2}I)^{-1}\mathbf{X}^t\mathbf{y} \end{aligned}$$

Setting the ridge regression parameter $\lambda = \frac{\sigma^2}{\tau^2}$, we see that the above solution is equivalent to ridge regression. Note how λ varies with σ^2 and τ^2. When τ^2 is small (meaning we have prior knowledge to believe β is close to zero), then λ is large (meaning large β will be penalized). When σ^2 is small (meaning that observations $\mathbf{y}$ are not noisy), we will focus on fitting the data (small λ).

In the same way that we showed that $p(\mu|\mathcal{X})$ is Gaussian when $p(x|\mu)$ and $p(\mu)$ are both Gaussian, you can show that $p(\beta|\mathbf{y}, \mathbf{X})$ is Gaussian. Using m_B and Σ_B for the mean and covariance of the posterior, then the exponent is:

$$-\frac{1}{2}(\beta - m_B)^t\Sigma_B^{-1}(\beta - m_B) = -\frac{1}{2}(\beta^t\Sigma_B^{-1}\beta - 2\beta^t\Sigma_B^{-1}m_B + m_B^t\Sigma_B^{-1}m_B)$$

Find Σ_B^{-1} by equating the second order beta terms from the $p(\beta|\mathbf{y}, \mathbf{X})$ equation above:

$$\beta^t\Sigma_B^{-1}\beta = \frac{1}{\sigma^2}\beta^t X^t X\beta + \frac{1}{\tau^2}\beta^t\beta = \frac{1}{\sigma^2}\beta^t(X^t X + \frac{\sigma^2}{\tau^2}I)\beta$$

which gives

$$\Sigma_B^{-1} = \frac{1}{\sigma^2}[X^t X + \frac{\sigma^2}{\tau^2}I]$$

Then you can show that the $\hat{\beta}$ estimate above is the mean by plugging in

$$m_B = (\mathbf{X}^t\mathbf{X} + \frac{\sigma^2}{\tau^2}I)^{-1}\mathbf{X}^t\mathbf{y} = \frac{1}{\sigma^2}\Sigma_B\mathbf{X}^t\mathbf{y}$$

$$\beta^t\Sigma_B^{-1}m_B = \frac{1}{\sigma^2}\beta^t\Sigma_B^{-1}\Sigma_B\mathbf{X}^t\mathbf{y} = \frac{1}{\sigma^2}\beta^t\mathbf{X}^t\mathbf{y}$$

which is the single β term in the $p(\beta|\mathbf{y}, \mathbf{X})$ equation. If $\hat{\beta}$ is the mean of the Gaussian, it is also the mode.

4. Implementation of Ridge Regression: The figures below illustrate the norm of β vs. λ for Ridge regression and the associated performance on training and test compared to least squares. As expected, the norm of β decreases as λ increases. The Least Squares solution is always better than the Ridge Regression *on the training data*, because the Least Squares solution minimizes this error (when λ is really small, the Ridge solution is essentially the same as the LS solution so they have the same error). However, for moderate values of λ the Ridge solution has lower error *on the test data* than the LS solution. At very large values of λ, both the training and test error are high for Ridge; lowest error on the test data occurs around $\lambda = 10$.

Figure 2: Norm of β vs. λ for Ridge regression

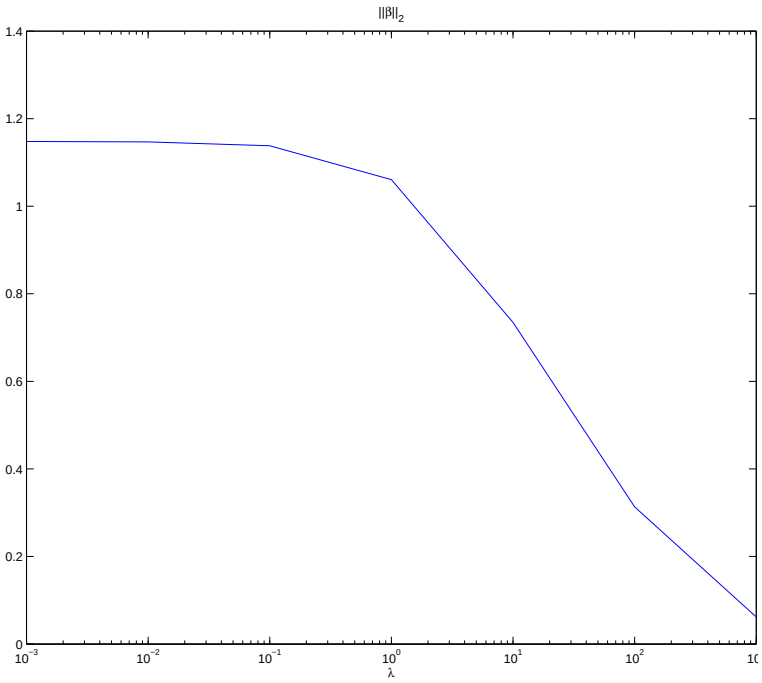


Figure 3: Errors on train and test for Ridge and Least Squares vs. λ