Selected Exercises of Elements of Statistical Learning

Mike Craig

Last Updated September 9, 2017

1 Introduction

This is a collection of certain exercise solutions I did as I was reading through this book. As I am not a mathematician, the ”proofs” listed here are rough and probably not rigorous. They only serve as a means to explain the concepts in a way that aids my own understanding.

2 Overview of Supervised Learning

Exercises

Exercise 2.1. **Suppose each of K-classes has an associated target t_k, which is a vector of all zeros, except a one in the kth position. Show that classifying to the largest element of $\hat{y}$ amounts to choosing the closest target, $min_k||t_k - \hat{y}||$, if the elements of $\hat{y}$ sum to one.**

We are trying to show that

$$arg\, max_k\, \hat{y_k} = arg\, min_k||t_k - \hat{y}||$$

To do this, we can show that for any $k^* \neq arg\, max_k\, \hat{y_k}$ and $k = arg\, max_k\, \hat{y_k}$,

$$||t_{k^*} - \hat{y}|| > ||t_k - \hat{y}||$$

Note that we can equivalently consider the $|| \cdot ||^2$ instead of $|| \cdot ||$, because they are both monotonic for ≥ 0.

Using the definition of the Euclidean norm,

$$\begin{align} ||t_{k^*} - \hat{y}||^2 &= ||\hat{y}||^2 + ||t_{k^*}||^2 - 2t_{k^*}\hat{y} \tag{1}\\ &= \hat{y}^2 \tag{2}\\ \tag{3} \end{align}$$

Similarly,

$$\begin{align} ||t_k - \hat{y}||^2 &= ||\hat{y}||^2 + ||t_k||^2 - 2t_k\hat{y} \tag{4}\\ &= \hat{y}^2 + 1 - 2\hat{y} \tag{5} \end{align}$$

Therefore

$$\begin{align} ||t_{k^*} - \hat{y}||^2 - ||t_k - \hat{y}||^2 &= \hat{y}^2 - (\hat{y}^2 + 1 - 2\hat{y}) \tag{6}\\ &= -1 + 2\hat{y} \tag{7}\\ &\geq 0 \tag{8} \end{align}$$

since $\hat{y}$ sums to one. So, $||t_k - \hat{y}||$ is minimized when $k = arg\, max_k\, \hat{y_k}$.

Exercise 2.2. **Show how to compute the Bayes decision boundary for the simulation example in Figure 2.5.**

Note that in the example in Figure 2.5, we have defined ($BLUE = 0$, $ORANGE = 1$), so this can be considered a binary choice, where the Bayes decision boundary can be defined by:

$$\hat{f}(X) = E(Y|X) = Pr(G = G_1|X) = Pr(ORANGE|X)$$

if G_1 corresponds to $Y = 1$. Since we know the structural model that generated this example, we can compute this. Note that

$$Pr(ORANGE|X) = \frac{1}{10}\sum_{i=0}^{10} g_i(x)$$

where $g_i(x) = N(\mu_i, I/5)$, where μ_i is one of the means generated during the generation process. Since these Gaussians are all i.i.d, we can write this mixture as:

$$\begin{aligned}
Pr(ORANGE|X) &= Pr(N(\mu_1 + \cdots + \mu_{10}, 10I/5) = X) && (9)\\
&= Pr(N(||\mu||, 2I) = X) && (10)\\
&= \frac{1}{(2\pi)^{p/2}}|\Sigma|^{1/2} exp\{-\frac{1}{2}(x - ||\mu||)^T\Sigma^{-1}(x - ||\mu||)\} && (11)\\
&= \frac{\sqrt{2}}{(2\pi)^{p/2}} exp\{-\frac{1}{4}(x - ||\mu||)^T(x - ||\mu||)\} && (12)\\
&= \frac{\sqrt{2}}{(2\pi)^{p/2}} exp\{-\frac{1}{4}(x^Tx - ||\mu||^2)\} && (13)
\end{aligned}$$

where μ is the mean of the Gaussians that make up the $ORANGE$ class. Now, we can plug in the Gaussians that make up the μ vector. Note that μ itself is a product of a Gaussian pdf. So, the distribution of $||\mu||$ is given by:

$$Pr(||\mu||) = \sum_{i=0}^{10} N((0,1)^T, I) = N((0,10)^T, 10I)$$

Plugging in the normal pdf above into the previous equation is tedious, but can be done and has a closed solution.

Exercise 2.3. **Consider N data points uniformly distributed in a p-dimensional unit ball centered at the origin. Suppose we consider a nearest-neighbor estimate at the origin. Show that the median distance from the origin to the closest data point is given by:** $d(p, N) = (1 - \frac{1}{2}^{1/N})^{1/p}$**.**

Let m be the median distance from the origin to the closest point. This means that the probability that all data points are further than m is 0.5. "Further" simply means a greater norm. Since samples x_i are i.i.d, we can more formally state this as:

$$\prod_{i=1}^{N} P(||x_i|| > m) = \frac{1}{2}$$

Note that we can flip this around to be $\prod_{i=1}^{N} P(||x_i|| \leq m) = \frac{1}{2}$. Now we can use the cumulative function of the uniform distribution as follows:

$$\begin{aligned}\prod_{i=1}^{N} P(||x_i|| \leq m) &= \prod_{i=1}^{N} 1 - ||m|| && (14)\\ &= \prod_{i=1}^{N} 1 - m^p && (15)\\ &= (1 - m^p)^N = \frac{1}{2} && (16)\end{aligned}$$

Now we can solve for m:

$$\begin{aligned}\frac{1}{2} &= (1 - m^p)^N && (17)\\ \frac{1}{2}^{1/N} &= 1 - m^p && (18)\\ m^p &= 1 - \frac{1}{2}^{1/N} && (19)\\ m &= (1 - \frac{1}{2}^{1/N})^{1/p} && (20)\end{aligned}$$

Exercise 2.4. **The edge effect problem discussed on page 23 is not peculiar to uniform sampling from bounded domains. Consider inputs drawn from a spherical multinormal distribution $XN(0, I_p)$. The squared distance from any sample point to the origin has a χ^2_p distribution with mean p. Consider a prediction point x_0 drawn from this distribution, and let $a = x_0/||x_0||$ be an associated unit vector. Let $z_i = a^T x_i$ be the projection of each of the training points on this direction. Show that the z_i are distributed $N(0, 1)$ with expected squared distance from the origin 1, while the target point has expected squared distance p from the origin. Hence for $p = 10$, a randomly drawn test point is about 3.1 standard deviations from the origin, while all the training points are on average one standard deviation along direction a. So most prediction points see themselves as lying on the edge of the training set.**

Exercise 2.5. **(a) Derive equation (2.27)**

(a) Equation (2.27) states

$$\begin{aligned}EPE(x_0) &= E_{y_0|x_0} E_\tau (y_0 - \hat{y}_0)^2 && (21)\\ &= Var(y_0|x_0) + E_\tau[\hat{y}_0 - E_\tau \hat{y}_0]^2 + [E_\tau \hat{y}_0 - x_0^T \beta]^2 && (22)\\ &= Var(y_0|x_0) + Var_\tau(\hat{y}_0) + Bias^2(\hat{y}_0) && (23)\\ &= \sigma^2 + E_\tau x_0^T (\mathbf{X}^T \mathbf{X})^{-1} x_0 \sigma^2 + 0^2. && (24)\end{aligned}$$

We can derive this is in the three constituent parts shown. Let's start with the bias squared term. Since we know that the least squares estimator is an unbiased estimator with respect to the training set, it is obvious that

$$[E_\tau \hat{y}_0 - x_0^T \beta]^2 = [x_0^T \beta - x_0^T \beta]^2 = 0^2$$

So, the bias term is zero. Next, we will look at the middle term. Note that for the least squares solution $\hat{\beta}$, $Var(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$. So,

$$\begin{aligned} Var_\tau(\hat{y}_0) &= Var_\tau(x_0^T \hat{\beta}) && (25)\\ &= x_0^2 Var(\hat{\beta}) && (26)\\ &= E_\tau x_0^T (\mathbf{X}^T\mathbf{X})^{-1} x_0 \sigma^2 && (27) \end{aligned}$$

Finally, it is clear that $Var(y_0|x_0) = \sigma^2$, since this is how we define σ^2.
(b) Equation (2.28) states

$$\begin{aligned} E_{x_0} EPE(x_0) &\sim E_{x_0} x_0^T Cov(X)^{-1} x_0 \sigma^2/N + \sigma^2 && (28)\\ &= trace[Cov(X)^{-1} Cov(x_0)]\sigma^2/N + \sigma^2 && (29)\\ &= \sigma^2(p/N) + \sigma^2 && (30) \end{aligned}$$

To derive this, we have to show that

$$E_{x_0} x_0^T Cov(X)^{-1} x_0 = trace[Cov(X)^{-1} Cov(x_0)]$$

To do this,

$$\begin{aligned} E_{x_0} x_0^T Cov(X)^{-1} x_0 &= E_{x_0} Cov(x_0^T X^{-1}, x_0^T X^{-1}) && (31)\\ &= E_{x_0} Cov(x_0) Cov(X)^{-1} && (32)\\ &= trace[Cov(x_0) Cov(X)^{-1}] && (33)\\ &= trace[Cov(X)^{-1} Cov(x_0)] && (34) \end{aligned}$$

The first line is evident through the linearity of the covariance matrix. Note that since the Xs are i.i.d, we can get from the first line to the second line. The third line is true because $Cov(x_0)$ is a scalar, and the fourth line is because of the cyclic property of traces.

Exercise 2.6. **Consider a regression problem with inputs x_i and outputs y_i, and a parameterized model $f_\theta(x)$ to be fit by least squares. Show that if there are observations with *tied* or *identical* values of x, then the fit can be obtained from a reduced weighted least squares problem.**

For a linear model fit with least squares, the loss function is defined as:

$$\begin{aligned} L(y,\hat{y}) &= E(y-\hat{y})^2 && (35)\\ &\approx \frac{1}{N}\sum_{i=0}^{N}(y_i-\hat{y}_i)^2 && (36)\\ &= \frac{1}{N}\sum_{i=0}^{N}(y_i-\beta^T X_i)^2 && (37) \end{aligned}$$

For a given model (i.e. conditioning on β), this quantity in the summation is a function of X (assuming no noise in y). Therefore, if we receive k instances of some value of X, this is equivalent to having k instances of this loss in the loss function. To show this, let k values be the same X_i:

$$\begin{aligned} L(y,\hat{y}) &= \frac{1}{k}\sum_{i=0}^{k}(y_i-\beta^T X_i)^2 + \frac{1}{N-k}\sum_{j=0;j\neq i}^{N-k}(y_j-\beta^T X_j)^2 && (38)\\ &= k(y_i-\beta^T X_i)^2 + \frac{1}{N-k}\sum_{j=0;j\neq i}^{N-k}(y_j-\beta^T X_j)^2 && (39) \end{aligned}$$

Therefore, this is equivalent to a weighted squares solution, where X_i is given a weight of k and every other point is given a weight of 1.