ELEMENTS OF STATISTICAL LEARNING - CHAPTER SOLUTIONS

ANDREW TULLOCH

1. Chapter 1

No exercises.

2. Chapter 2

Exercise 2.1. *Suppose that each of K-classes has an associated target t_k, which is a vector of all zeroes, except a one in the k-th position. Show that classifying 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.*

Proof. The assertion is equivalent to showing that

$$\arg\max_i \hat{y}_i = \arg\min_k \|t_k - \hat{y}\| = \arg\min_k \|\hat{y} - t_k\|^2$$

by monotonicity of $x \mapsto x^2$ and symmetry of the norm.

WLOG, let $\|\cdot\|$ be the Euclidean norm $\|\cdot\|_2$. Let $k = \arg\max_i \hat{y}_i$, with $\hat{y}_k = \max y_i$. Note that then $\hat{y}_k \geq \frac{1}{K}$, since $\sum \hat{y}_i = 1$.

Then for any $k' \neq k$ (note that $y_{k'} \leq y_k$), we have

$$\begin{aligned} \|y - t_{k'}\|_2^2 - \|y - t_k\|_2^2 &= y_k^2 + (y_{k'} - 1)^2 - \left(y_{k'}^2 + (y_k - 1)^2\right) \\ &= 2\left(y_k - y_{k'}\right) \\ &\geq 0 \end{aligned}$$

since $y_{k'} \leq y_k$ by assumption.

Thus we must have

$$\arg\min_k \|t_k - \hat{y}\| = \arg\max_i \hat{y}_i$$

as required. □

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

Proof. The Bayes classifier is

$$\hat{G}(X) = \arg\max_{g \in \mathcal{G}} P(g|X = x).$$

In our two-class example ORANGE and BLUE, the decision boundary is the set where

$$P(g = \text{BLUE}|X = x) = P(g = \text{ORANGE}|X = x) = \frac{1}{2}.$$

By the Bayes rule, this is equivalent to the set of points where

$$P(X = x|g = \text{BLUE})P(g = \text{BLUE}) = P(X = x|g = \text{ORANGE})P(g = \text{ORANGE})$$

And since we know $P(g)$ and $P(X = x|g)$, the decision boundary can be calculated. □

Exercise 2.3. *Consider N data points uniformly distributed in a p-dimensional unit ball centered at the origin. Show the the median distance from the origin to the closest data point is given by*

$$d(p, N) = \left(1 - \left(\frac{1}{2}\right)^{1/N}\right)^{1/p}$$

Proof. Let r be the median distance from the origin to the closest data point. Then

$$P(\text{All } N \text{ points are further than } r \text{ from the origin}) = \frac{1}{2}$$

by definition of the median.

Since the points x_i are independently distributed, this implies that

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

and as the points x_i are uniformly distributed in the unit ball, we have that

$$\begin{aligned} P(\|x_i\| > r) &= 1 - P(\|x_i\| \leq r) \\ &= 1 - \frac{Kr^p}{K} \\ &= 1 - r^p \end{aligned}$$

Putting these together, we obtain that

$$\frac{1}{2} = (1 - r^p)^N$$

and solving for r, we have

$$r = \left(1 - \left(\frac{1}{2}\right)^{1/N}\right)^{1/p}$$

□

Exercise 2.4. *Consider inputs drawn from a spherical multivariate-normal distribution $X \sim N(0, \mathbf{1}_p)$. 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 $a = \frac{x_0}{\|x_0\|}$*

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.

Proof. Let $z_i = a^T x_i = \frac{x_0^T}{\|x_0\|} x_i$. Then z_i is a linear combination of $N(0,1)$ random variables, and hence normal, with expectation zero and variance

$$\mathrm{Var}(z_i) = \|a^T\|^2 \mathrm{Var}(x_i) = \mathrm{Var}(x_i) = 1$$

as the vector a has unit length and $x_i \sim N(0,1)$.

For each target point x_i, the squared distance from the origin is a χ_p^2 distribution with mean p, as required. □

Exercise 2.5. *(a) Derive equation (2.27) in the notes.*
(b) Derive equation (2.28) in the notes.

Proof. (i) We have

$$\begin{aligned} EPE(x_0) &= E_{y_0|x_0} E_{\mathcal{T}} (y_0 - \hat{y}_0)^2 \\ &= \mathrm{Var}(y_0|x_0) + E_{\mathcal{T}}[\hat{y}_0 - E_{\mathcal{T}}\hat{y}_0]^2 + [E_{\mathcal{T}} - x_0^T\beta]^2 \\ &= \mathrm{Var}(y_0|x_0) + \mathrm{Var}_{\mathcal{T}}(\hat{y}_0) + \mathrm{Bias}^2(\hat{y}_0). \end{aligned}$$

We now treat each term individually. Since the estimator is unbiased, we have that the third term is zero. Since $y_0 = x_0^T\beta + \epsilon$ with ϵ an $N(0, \sigma^2)$ random variable, we must have $\mathrm{Var}(y_0|x_0) = \sigma^2$.

The middle term is more difficult. First, note that we have

$$\begin{aligned} \mathrm{Var}_{\mathcal{T}}(\hat{y}_0) &= \mathrm{Var}_{\mathcal{T}}(x_0^T\hat{\beta}) \\ &= x_0^T \mathrm{Var}_{\mathcal{T}}(\hat{\beta}) x_0 \\ &= E_{\mathcal{T}} x_0^T \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1} x_0 \end{aligned}$$

by conditioning (3.8) on $\mathcal{T}$.

(ii)

□

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 with 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.*

Proof. This is relatively simple. WLOG, assume that $x_1 = x_2$, and all other observations are unique. Then our RSS function in the general least-squares estimation is

$$RSS(\theta) = \sum_{i=1}^{N} (y_i - f_\theta(x_i))^2 = \sum_{i=2}^{N} w_i (y_i - f_\theta(x_i))^2$$

where

$$w_i = \begin{cases} 2 & i = 2 \\ 1 & \text{otherwise} \end{cases}$$

Thus we have converted our least squares estimation into a reduced weighted least squares estimation. This minimal example can be easily generalised. □

Exercise 2.7. *Suppose that we have a sample of N pairs x_i, y_i, drawn IID from the distribution such that $x_i \sim h(x), y_i = f(x_i) + \epsilon_i, E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$.*

We construct an estimator for f linear in the y_i,

$$\hat{f}(x_0) = \sum_{i=1}^{N} \ell_i(x_0; \mathcal{X}) y_i$$

where the weights $\ell_i(x_0; X)$ do not depend on the y_i, but do depend on the training sequence x_i denoted by $\mathcal{X}$.

(a) Show that the linear regression and k-nearest-neighbour regression are members of this class of estimators. Describe explicitly the weights $\ell_i(x_0; \mathcal{X})$ in each of these cases.

Proof. (a) Recall that the estimator for f in the linear regression case is given by

$$\hat{f}(x_0) = x_0^T \beta$$

where $\beta = (X^T X)^{-1} X^T y$. Then we can simply write

$$\hat{f}(x_0) = \sum_{i=1}^{N} \left(x_0^T (X^T X)^{-1} X^T \right)_i y_i.$$

Hence

$$\ell_i(x_0; \mathcal{X}) = \left(x_0^T (X^T X)^{-1} X^T \right)_i.$$

In the k-nearest-neighbour representation, we have

$$\hat{f}(x_0) = \sum_{i=1}^{N} \frac{y_i}{k} \mathbf{1}_{x_i \in N_k(x_0)}$$

where $N_k(x_0)$ represents the set of k-nearest-neighbours of x_0. Clearly,

$$\ell_i(x_0; \mathcal{X}) = \frac{1}{k} \mathbf{1}_{x_i \in N_k(x_0)}$$

□