第二次作业

Ex 2.1

Ex. 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.

I don't know what the problem is trying to say after half a dozen readings. I can only solve the question as I understand it.

We aim to prove that:

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

Notice that $y = (\|x\|)^2$ is monotonic in the positive range, so we can easily see that:

$$\arg\min_k \|t_k - \hat{y} = \arg\min_k (\|t_k - \hat{y}\|)^2 = \arg\min_k (t_k - \hat{y})^\top (t_k - \hat{y}) = \arg\min_k (t_k^\top t_k - 2t_k^\top \hat{y} + \hat{y}^\top \hat{y})$$

By the definition of t_k and $\hat{y}$, we know that $t_k^\top t_k = 1$ and $\hat{y}^\top \hat{y}$ is independent of k.

So,

$$\arg\min_k (t_k^\top t_k - 2t_k^\top \hat{y} + \hat{y}^\top \hat{y}) = \arg\max_k (t_k^\top \hat{y}) = \arg\max_k \hat{y}_k$$

Ex 2.3

Ex. 2.3 Derive equation (2.24).

Another consequence of the sparse sampling in high dimensions is that all sample points are close to an edge of the sample. 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. The median

distance from the origin to the closest data point is given by the expression

$$d(p,N) = \left(1 - \frac{1}{2}^{1/N}\right)^{1/p} \tag{2.24}$$

This Ex. 2.3 was actually done in the first assignment.

Since the probabilities are uniformly distributed, the probability of landing in a particular region is equal to the ratio of that region to the total region.

Thus we get the probability that a data point is no more than r away from the origin point:

$$\frac{r^p}{1^p} = r^p$$

Thus it is easy to obtain that the probability that the distance of the nearest data point to the origin is greater than r for N randomly chosen points is

$$(1 - r^p)^N$$

Finding the median distance makes the above equation $\frac{1}{2}$, so it can be solved:

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

Ex 2.4

Ex. 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 $X \sim N(0, \mathbf{I}_p)$. 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.

Notice that:

$$z_i = a^\top x_i = \frac{x_0^\top x_i}{\|x_0\|}$$

and we define $\overline{x}_i = \frac{x_i}{\|x_i\|}$, so $z_i = \frac{x_0^\top \overline{x}_i}{\|x_0\|} \cdot \|x_i\|$, it's easy to see that $\|\overline{x}_i\| = \sum_{j=1}^{p} (\overline{x}_i[j])^2 = 1$

Considering the symmetry of the spherical shape.

Without loss of generality, we can transform the coordinate system such that $a = (1, 0, \dots, 0)^\top$, we can find the first half:

$$E[(\frac{x_0^\top \overline{x}_i}{\|x_0\|})^2] = E[(\overline{x}_i[1])^2] = \frac{1}{p}\sum_{j=1}^{p}(\overline{x}_i[j])^2 = \frac{1}{p}$$

According to the description in the exercise, there are $E[\|x_i\|^2] = p$, so

$$E\|z_i\|^2 = E[\|\frac{x_0^\top \overline{x}_i}{\|x_0\|} \cdot \|x_i\|\|^2] = E[(\frac{x_0^\top \overline{x}_i}{\|x_0\|})^2] \cdot E[\|x_i\|^2] = \frac{p}{p} = 1$$

Noting that there is $\sqrt{p} = \sqrt{10} \approx 3.16$, then it clearly satisfies.

Ex 2.7

Ex. 2.7 Suppose we have a sample of N pairs x_i, y_i drawn i.i.d. from the distribution characterized as follows:

$$x_i \sim h(x), \text{ the design density}$$
$$y_i = f(x_i) + \varepsilon_i, \ f \text{ is the regression function}$$
$$\varepsilon_i \sim (0, \sigma^2) \text{ (mean zero, variance } \sigma^2)$$

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

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

(b) Decompose the conditional mean-squared error

$$\mathrm{E}_{\mathcal{Y}|\mathcal{X}}(f(x_0) - \hat{f}(x_0))^2$$

into a conditional squared bias and a conditional variance component. Like $\mathcal{X}$, $\mathcal{Y}$ represents the entire training sequence of y_i.

(c) Decompose the (unconditional) mean-squared error

$$\mathrm{E}_{\mathcal{Y},\mathcal{X}}(f(x_0) - \hat{f}(x_0))^2$$

into a squared bias and a variance component.

(d) Establish a relationship between the squared biases and variances in the above two cases.

(a)

Linear Regression

Following the calculation of section 2.3.1, to reduce the RSS, we should let

$$f(x_0) = [1; x_0]\hat{\beta} = [1; x_0](\mathcal{X}^\top \mathcal{X})^{-1} \mathcal{X}^\top y$$

To simplify, we include a constant variable 1 in $x_i \in \mathcal{X}$ but not x_0.

So we follow the inner product expension,

$$\ell_i(x_0;\mathcal{X}) = [1;x_0](\mathcal{X}^\top\mathcal{X})^{-1}{x_i}^\top$$

k-nearest-neighbor Regression

Define $N_k(x)$ is the neighborhood of x defined by the k cloest points $x_i \in \mathcal{X}$

So

$$\ell_i(x_0;\mathcal{X}) = \begin{cases} \frac{1}{k}, & x_i \in N_k(x_0) \\ 0, & Else \end{cases}$$

(b)

Just expand the square terms and then combine like terms

$$\begin{aligned} \mathrm{E}_{\mathcal{Y}|\mathcal{X}}\left(\left(f(x_0)-\hat{f}(x_0)\right)^2\right) &= f(x_0)^2 - 2\cdot f(x_0)\cdot \mathrm{E}_{\mathcal{Y}|\mathcal{X}}\left(\hat{f}(x_0)\right) + \mathrm{E}_{\mathcal{Y}|\mathcal{X}}\left(\left(\hat{f}(x_0)\right)^2\right) \\ &= \left(f(x_0)-\mathrm{E}_{\mathcal{Y}|\mathcal{X}}\left(\hat{f}(x_0)\right)\right)^2 + \mathrm{E}_{\mathcal{Y}|\mathcal{X}}\left(\left(\hat{f}(x_0)\right)^2\right) - \left(\mathrm{E}_{\mathcal{Y}|\mathcal{X}}\left(\hat{f}(x_0)\right)\right)^2 \\ &= (\mathrm{bias})^2 + \mathrm{Var}(\hat{f}(x_0)) \end{aligned}$$

(c)

It's basically the same as the question above.

$$\begin{aligned} \mathrm{E}_{\mathcal{X},\mathcal{Y}}\left(\left(f(x_0)-\hat{f}(x_0)\right)^2\right) &= f(x_0)^2 - 2\cdot f(x_0)\cdot \mathrm{E}_{\mathcal{X},\mathcal{Y}}\left(\hat{f}(x_0)\right) + \mathrm{E}_{\mathcal{X},\mathcal{Y}}\left(\left(\hat{f}(x_0)\right)^2\right) \\ &= \left(f(x_0)-\mathrm{E}_{\mathcal{X},\mathcal{Y}}\left(\hat{f}(x_0)\right)\right)^2 + \mathrm{E}_{\mathcal{X},\mathcal{Y}}\left(\left(\hat{f}(x_0)\right)^2\right) - \left(\mathrm{E}_{\mathcal{X},\mathcal{Y}}\left(\hat{f}(x_0)\right)\right)^2 \\ &= (\mathrm{bias})^2 + \mathrm{Var}(\hat{f}(x_0)) \end{aligned}$$

(d.b)

$$\mathrm{bias} = f(x_0) - \mathrm{E}_{\mathcal{Y}|\mathcal{X}}\left(\hat{f}(x_0)\right) = f(x_0) - \sum_{i=1}^{N} \ell_i(x_0,\mathcal{X})f(x_i)$$

$$
\mathrm{Var}(\hat{f}(x_0)) = \mathrm{E}_{\mathcal{Y}|\mathcal{X}}\left(\left(\hat{f}(x_0)\right)^2\right) - \left(\mathrm{E}_{\mathcal{Y}|\mathcal{X}}\left(\hat{f}(x_0)\right)\right)^2
$$

$$
= \left(\sum_{i,j} \ell_i(x_0;\mathcal{X})\ell_j(x_0;\mathcal{X})f(x_i)f(x_j) + \sum_i \sigma^2 \ell_i(x_0;\mathcal{X})^2\right) - \left(\sum_{i=1}^N \ell_i(x_0,\mathcal{X})f(x_i)\right)^2
$$

(d.c)

Let $\mathcal{X} = (x_1, \ldots, x_n)$, and we define $d\mathcal{X} = h(x_1)\ldots h(x_n)dx_1\ldots dx_n$

$$
\text{bias} = f(x_0) - \mathrm{E}_{\mathcal{X},\mathcal{Y}}\left(\hat{f}(x_0)\right) = f(x_0) - \int \ell_i(x_0,\mathcal{X})f(x_i)d\mathcal{X}
$$

$$
\mathrm{Var}(\hat{f}(x_0)) = \mathrm{E}_{\mathcal{X},\mathcal{Y}}\left(\left(\hat{f}(x_0)\right)^2\right) - \left(\mathrm{E}_{\mathcal{X},\mathcal{Y}}\left(\hat{f}(x_0)\right)\right)^2
$$

$$
= \left(\int \ell_i(x_0;\mathcal{X})\ell_j(x_0;\mathcal{X})f(x_i)f(x_j)d\mathcal{X} + \int \sigma^2 \ell_i(x_0;\mathcal{X})^2 d\mathcal{X}\right) - \left(\int \ell_i(x_0,\mathcal{X})f(x_i)d\mathcal{X}\right)^2
$$

Ex. 3.5

Ex. 3.5 Consider the ridge regression problem (3.41). Show that this problem is equivalent to the problem

$$
\hat{\beta}^c = \underset{\beta^c}{\operatorname{argmin}}\left\{\sum_{i=1}^N\left[y_i - \beta_0^c - \sum_{j=1}^p (x_{ij} - \bar{x}_j)\beta_j^c\right]^2 + \lambda\sum_{j=1}^p \beta_j^{c2}\right\}. \tag{3.85}
$$

Give the correspondence between β^c and the original β in (3.41). Characterize the solution to this modified criterion. Show that a similar result holds for the lasso.

Notice that the difference between $\hat{\beta}^c$ and β is:

$$
\hat{\beta}^c = \underset{\beta^c}{\operatorname{argmin}}\left\{\sum_{i=1}^N\left[y_i - \beta_0^c - \sum_{j=1}^p (x_{ij} - \bar{x}_j)\beta_j^c\right]^2 + \lambda\sum_{j=1}^p \beta_j^{c2}\right\}
$$

$$
= \underset{\beta^c}{\operatorname{argmin}}\left\{\sum_{i=1}^N\left[y_i - \beta_0^c + \sum_{j=1}^p \bar{x}_j\beta_j^c - \sum_{j=1}^p x_{ij}\beta_j^c\right]^2 + \lambda\sum_{j=1}^p \beta_j^{c2}\right\}
$$

If we define $\beta_0 = \beta_0^c - \sum_{j=1}^p \bar{x}_j \beta_j^c$, then $\hat{\beta}^c$ and β are always the same.

Ex. 3.6

> Ex. 3.6 Show that the ridge regression estimate is the mean (and mode) of the posterior distribution, under a Gaussian prior $\beta \sim N(0, \tau\mathbf{I})$, and Gaussian sampling model $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$. Find the relationship between the regularization parameter λ in the ridge formula, and the variances τ and σ^2.

By using Bayes's theorem and taking logarithm on both side, we get

$$\begin{aligned} P(\beta|\mathbf{y}) &= \frac{P(\mathbf{y}|\beta)P(\beta)}{P(\mathbf{y})} \\ &= \frac{N(\mathbf{X}\beta, \sigma^2\mathbf{I})N(0, \tau\mathbf{I})}{P(\mathbf{y})}. \end{aligned}$$

and

$$\ln(P(\beta|\mathbf{y})) = -\frac{1}{2}\left(\frac{(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)}{\sigma^2} + \frac{\beta^T\beta}{\tau}\right) + C,$$

because $P(\mathbf{y})$ is independent of β. Then we can just maximazing

$$\arg\max_\beta \left((\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta) + \frac{\sigma^2\beta^T\beta}{\tau}\right)$$

and we can define $\lambda = \frac{\sigma^2}{\tau}$

Ex. 3.7

> Ex. 3.7 Assume $y_i \sim N(\beta_0 + x_i^T\beta, \sigma^2), i = 1, 2, \ldots, N$, and the parameters $\beta_j,\ j = 1, \ldots, p$ are each distributed as $N(0, \tau^2)$, independently of one another. Assuming σ^2 and τ^2 are known, and β_0 is not governed by a prior (or has a flat improper prior), show that the (minus) log-posterior density of β is proportional to $\sum_{i=1}^N (y_i - \beta_0 - \sum_j x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^p \beta_j^2$ where $\lambda = \sigma^2/\tau^2$.

Using Bayes theorem,

$$P(\beta|\mathbf{y}) = \frac{P(\mathbf{y}|\beta)P(\beta)}{P(\mathbf{y})}.$$

And we know that (C_1, C_2 substitute two constants that we don't care about)

$$P(\beta) = C_1 \exp\left(-\frac{\left\|\beta\right\|^2}{2\tau^2}\right)$$
$$P(\mathbf{y}|\beta) = C_2 \exp\left(-\frac{\left\|\mathbf{y} - X\beta\right\|^2}{2\sigma^2}\right)$$

So

$$-\ln(P(\beta|\mathbf{y})) = \frac{1}{2\sigma^2}\left(\left\|\mathbf{y} - X\beta\right\|^2 + \frac{\sigma^2}{\tau^2}\left\|\beta\right\|^2\right) + C,$$

and we can let $\lambda = \frac{\sigma^2}{\tau^2}$. And the claim holds iff $C = 0$