## Suggestions - Problem Set 1

**Hastie, 2.5 (a)** This will be similar to that done in class, with use of equation (3.8) in the last line.

Some comments about notation -- see the Notes on matrix notation on the web page for more details. If $\mathbf{y} = (y_0, y_1, ..., y_p)^T$ is a random vector, then the expression $V(\mathbf{y})$ is the corresponding *covariance matrix*, with $i, j$ component $(V(\mathbf{y}))_{ij} = E[y_i - \overline{y}_i)(y_j - \overline{y}_j)]$, where in general we denote $\overline{a}_i = E(a_i)$ as the mean value of $a_i$. However, if $y$ is a scalar (non-vector) random variable, then the same notation $V(y) = E[(y - \overline{y})^2]$ represents the variance of $y$ (now a single number).

A comment about eq. (3.8). We are computing $V(\widehat{\beta})$, the *covariance matrix* of the random vector $\widehat{\beta} = (\widehat{\beta}_0, ..., \widehat{\beta}_p)^T$. The $i, j$ entry of this matrix is

$$V(\widehat{\beta})_{ij} = E[\widehat{\beta}_i - \overline{\widehat{\beta}}_i][\widehat{\beta}_j - \overline{\widehat{\beta}}_j].$$

In (3.8), note we are *assuming* we know the $x$ part of the dataset, i.e., the matrix $\mathbf{X}$, but *not* the $\mathbf{Y}$ part. We are taking the expectation with respect to $\mathbf{y}$ but not $\mathbf{X}$. Thus appropriate subscripts here would be

$$V(\widehat{\beta}) = V_{\mathbf{y}|\mathbf{X}}(\widehat{\beta}) = E_{\mathbf{y}|\mathbf{X}}(\widehat{\beta} - E_{\mathbf{y}|\mathbf{X}}(\widehat{\beta}))^2.$$

We are also given

$$V_{y|\mathbf{X}}(\widehat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}; \tag{2}$$

note $\sigma^2$ is just the constant (non-matrix) variance of the error $\epsilon$, while $(\mathbf{X}^T\mathbf{X})^{-1}$ is now a fixed matrix. We are treating $\mathbf{y} = (y_1, ..., y_N)^T$ as a random variable $-$ this is why $V(\widehat{\beta})$ contains a $\sigma$ term in (2) above.

In (2.27), we no longer treat $\mathbf{X} \equiv \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix}$ as a fixed matrix - we treat each data point $\mathbf{x}_i$ in $\mathcal{T}$ as a random variable with some unknown but fixed distribution $p(\mathbf{x})$, which is why we take expectations over $\mathbf{X}$ below. We have (why is $\mathbf{x}_0$ independent of $\mathbf{X}$?):

$$\text{Var}_{\mathcal{T}}(\widehat{y}_0) = V_{\mathcal{T}}(\widehat{y}_0) = V_{\mathcal{T}}(\mathbf{x}_0{}^T\widehat{\beta}) = \mathbf{x}_0 V_{\mathcal{T}}(\widehat{\beta})\mathbf{x}_0^T .$$

Show (you may use (3.8))

$$V_{\mathcal{T}}(\widehat{\beta}) = E_{\mathcal{T}}[(\widehat{\beta} - \overline{\widehat{\beta}})^2] = E_{\mathbf{X},\mathbf{y}}[(\widehat{\beta} - \overline{\widehat{\beta}})^2]$$

$$= E_{\mathbf{X}} E_{\mathbf{y}|\mathbf{X}}[(\widehat{\beta} - \overline{\widehat{\beta}})^2] = \sigma^2 E_{\mathbf{X}}[(\mathbf{X}^T\mathbf{X})^{-1}].$$

Thus show

$$V_{\mathcal{T}}(\widehat{y}_0) = \sigma^2 \mathbf{x}_0 E_{\mathbf{X}}[(\mathbf{X}^T\mathbf{X})^{-1}]\mathbf{x}_0^T$$

and

$$\text{EPE}(\mathbf{x}_0) = E_{y_0|\mathbf{x}_0} E_{\mathcal{T}}(y_0 - \widehat{y}_0)^2$$

$$= V(y_0|\mathbf{x}_0) + E_{\tau}[\widehat{y}_0 - E_{\tau}(\widehat{y}_0)]^2 + [E_{\mathcal{T}}\widehat{y}_0 - \mathbf{x}_0^T \beta]^2$$

$$= V(y_0|\mathbf{x}_0) + V_{\mathcal{T}}(\widehat{y}_0) + \text{Bias}^2(\widehat{y}_0)$$

$$= V(y_0|\mathbf{x}_0) + \sigma^2 \mathbf{x}_0 E_{\mathbf{X}}[(\mathbf{X}^T\mathbf{X})^{-1}]\mathbf{x}_0^T + \text{Bias}^2(\widehat{y}_0).$$

Why is this the same as (2.27)? Now note $\mathcal{T} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ contains all of the information in **X** and there is no **y** in the expectation. Why can we replace $E_{\mathbf{X}}[(\mathbf{X}^T\mathbf{X})^{-1}]$ by $E_{\mathcal{T}}[(\mathbf{X}^T\mathbf{X})^{-1}]$ above?

Final remark: if you compare equation (2.27) with the analogous equation in the class notes, notice the last two squared terms appear in a different order in (2.27). But equation (2.27) is exactly the equation in the notes, specialized to the case of linear regression.

**(b)** We need to compute

$$E_{\mathbf{x}_0}\mathbf{x}_0^T \, \text{Cov}(X)^{-1}\mathbf{x}_0.$$

Note that $X = (X_0, ..., X_p)^T$ is the random vector giving the underlying distribution of the coordinates of a typical input data point $\mathbf{x} = (x_1, \ldots, x_p)$ (i.e., **x** is one of the points in the training set $\mathcal{T}$), and $\text{Cov}(X)$ is the covariance matrix, i.e., $\text{Cov}(X)_{ij} = E[(X_i - E(X_i)(X_j - E(X_j))]$. Letting $W = \text{Cov}(X)^{-1}$, show

$$E_{\mathbf{x}_0}\mathbf{x}_0^T W\mathbf{x}_0 = E_{\mathbf{x}_0}\text{tr}\big[\mathbf{x}_0^T W\mathbf{x}_0\big] = E_{\mathbf{x}_0}\text{tr}\big[W\mathbf{x}_0\mathbf{x}_0^T\big] = \text{tr}\big[W E_{\mathbf{x}_0}\big(\mathbf{x}_0\mathbf{x}_0^T\big)\big]$$

$$\text{tr}[W\text{Cov}(\mathbf{x}_0)] = \text{tr}\big[(\text{Cov } X)^{-1}\text{Cov}(\mathbf{x}_0)\big] = p. \tag{3}$$

Why were we allowed to add in the trace above? How was $\text{tr}(AB) = \text{tr}(BA)$ used?. Why are the random vectors $X$ and $\mathbf{x}_0$ (also viewed as random) identically distributed? Hence verify the last equality in (3) above.

**Hastie, Problem 2.7 (a)** For the case of linear regression, show

$$\widehat{f}(x_0) = \widehat{y}_0 = \mathbf{x}_0^T\widehat{\beta} = J^T\mathbf{y},$$

where $J^T = \mathbf{x}_0(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}$. What are the dimensions of $J^T$? Thus show we can write

$$\widehat{f}(\mathbf{x}_0) = \sum_{i=1} J_i y_i.$$

For $k$-nearest neighbors, show

$$\widehat{f}(\mathbf{x}_0) = \frac{1}{k} \sum_{\mathbf{x}_i \in N_k(\mathbf{x})} y_i \ ,$$

where $N_k(\mathbf{x})$ is the collection of $k$ nearest neighbors in the set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$. Show

$$\widehat{f}(\mathbf{x}_0) = \frac{1}{k} \sum_{i=1}^N \ell_i(\mathbf{x}_0, \mathcal{X}) y_i \ .$$

where

$$\ell_i(\mathbf{x}_0, \mathcal{X}) = \begin{cases} 1 & \text{if } \mathbf{x}_i \in N_k(\mathbf{x}_0) \\ 0 & \text{otherwise} \end{cases} .$$

**(b)** Justify that

$$E_{Y|X}(f(\mathbf{x}_0) - \widehat{f}(\mathbf{x}_0))^2 = E_{Y|X}(f(\mathbf{x}_0) - E_{Y|X}\widehat{f}(\mathbf{x}_0) + E_{Y|X}\widehat{f}(\mathbf{x}_0) - \widehat{f}(\mathbf{x}_0))^2$$

$$= (f(\mathbf{x}_0) - E_{Y|X}\widehat{f}(\mathbf{x}_0))^2 + E_{Y|X}(\widehat{f}(\mathbf{x}_0) - E_{Y|X}\widehat{f}(\mathbf{x}_0))^2$$

notice that the first part of the last expression is fixed and not a random variable.

$$= \text{bias}^2 + \text{Variance}$$

**(c)** Show we have exactly the same expression as above:

$$E_{Y,X}(f(\mathbf{x}_0) - \widehat{f}(\mathbf{x}_0))^2 = (f(\mathbf{x}_0) - E_{Y,X}\widehat{f}(\mathbf{x}_0))^2 + E_{Y,X}(\widehat{f}(\mathbf{x}_0) - E_{Y,X}\widehat{f}(\mathbf{x}_0))^2$$

What are the parts?

**(d)** For any random variable $A(\mathbf{X},\mathbf{Y})$ depending on random vectors **X** and **Y,**

$$E_{Y,X}(A(\mathbf{X},Y)) = E_{\mathbf{X}}[E_{Y|\mathbf{X}}(A(\mathbf{X},\mathbf{Y})].$$

Thus show

$$E_{Y,\mathbf{X}}(f(\mathbf{x}_0) - \widehat{f}(\mathbf{x}_0))^2 = E_{\mathbf{X}}E_{Y|\mathbf{X}}(f(\mathbf{x}_0) - \widehat{f}(\mathbf{x}_0))^2$$

$$= E_{\mathbf{X}}(f(\mathbf{x}_0) - E_{Y|\mathbf{X}}\widehat{f}(\mathbf{x}_0))^2 + E_{\mathbf{X}}E_{Y|\mathbf{X}}(\widehat{f}(\mathbf{x}_0) - E_{Y|\mathbf{X}}\widehat{f}(\mathbf{x}_0))^2$$

This gives an alternative to the expression for the same error in **(c)** - try to comment on it.