1. Describe the main principles of statistical learning in short and concise words using each of these terms.

Solution.
Statistical Learning refers to problem solving framework involving vast amount real-world data and statistical theory principles. Some of the modern day examples include email spam classification and stock market prediction. In short, we can call it as learning from data.

Broadly, these learning problems can be classified as either **supervised learning** – learning process guided by input and **outcome variables** or **unsupervised learning** – learning methodology where we observe only **input features** and without any measurements of responses. Based on the input data, the supervised learning methods can be of two types: **classification** - problem of identifying to which of a set of a category a new data belongs or **regression** – estimating a function from the inputs and numeric/ continuous output variables. In the context of classification tasks, we usually think of target variables as **qualitative** (Dogs vs. Cats classification) usually categorical labels and in regression analysis, outputs are usually **quantitative variables** i.e., numerical values.

The general learning framework for statistical models involve creating data in the form of **training data**, which has both input features and outcomes and **test data**, which has access only the inputs variables. As the name suggests, training data is used to train a model to learn the underlying data distribution. The trained model is usually used for either **prediction** tasks that involve predicting outcomes for new data points (such as test data) or for **inference** which primarily involves understanding how a stochastic model generates its estimates.

The process of building statistical models usually involve making different assumptions to approximate the underlying function. Usually, **parametric models** assume that a finite set parameters define/control the data distribution whereas **non-parametric models** assume that the underlying data cannot assumed with a finite set of parameters. Good examples for parametric and non-parametric models include linear regression models and k-nearest neighbour method respectively. ■

Problem 2 :

Let Y be a random variable. Show that

$$E(Y) = \text{argmin}_c E\left[(Y-c)^2\right]$$

Solution :

$$E(Y) = \text{argmin}_c E\left[(Y-c)^2\right]$$

L.H.S. ⇔ R.H.S.

→ Consider R.H.S, Let $c^* = \text{argmin}_c E\left[(Y-c)^2\right]$

→ $(y-c)^2$ is a convex function with a single minimum (Global minimum) with parameter c

→ c^* can be found by equating the first derivative with respect to $c \to 0$

$$\Rightarrow \frac{\partial\left(E\left[(Y-c)^2\right]\right)}{\partial c} = 0$$

$\{(a-b)^2 = a^2 + b^2 - 2ab$

$$\Rightarrow \frac{\partial\left(E\left[Y^2 + c^2 - 2Yc\right]\right)}{\partial c} = 0$$

$$\Rightarrow \frac{\partial\left(E[Y^2] + E[c^2] - 2E[Yc]\right)}{\partial c} = 0$$

$$\Rightarrow \frac{\partial E[Y^2]}{\partial c} + \frac{\partial E[c^2]}{\partial c} - \frac{\partial\left(2E[Yc]\right)}{\partial c} = 0$$

$$\Rightarrow 0 + 2c - 2E[Y] = 0$$

$$\Rightarrow c = E[Y] = \text{L.H.S}$$

So $\boxed{E[Y] = \text{argmin}_c E\left[(Y-c)^2\right]}$ Proof

PRIYANKA UPADHYAY

Problem 3

Prove that Bias-Variance tradeoff with irreducible error. Please Note that you should prove both equalities.

$$E\left[\{y_0 - \hat{f}(x_0)\}^2\right]$$

$$= E\left[\{\hat{f}(x_0) - E(\hat{f}(x_0))\}^2\right] + \left[E\{\hat{f}(x_0)\} - f(x_0)\right]^2 + \text{Var}(\varepsilon)$$

$$= \text{Var}(\hat{f}(x_0)) + \left[\text{Bias}(\hat{f}(x_0))\right]^2 + \text{Var}(\varepsilon)$$

Solution

As we know $y_0 = f(x_0) + \varepsilon$

where ε is random number with expected value $\hat{\varepsilon} = E[\varepsilon] = 0$ and variance $E\left[(\varepsilon - \hat{\varepsilon})^2\right] = E[\varepsilon^2] = \sigma^2 = \text{Var}(\varepsilon)$

So,

$$E\left[\{y_0 - \hat{f}(x_0)\}^2\right] \quad -①$$

$$\Rightarrow E\left[\{\underbrace{f(x_0) + \varepsilon}_{y_0 \text{ value}} - \hat{f}(x_0)\}^2\right]$$

$$\Rightarrow E\left[\{\underbrace{f(x_0) - \hat{f}(x_0)}_{a} + \underbrace{\varepsilon}_{b}\}^2\right] \qquad \because (a+b)^2 = a^2 + b^2 + 2ab$$

$$\Rightarrow E\left[\{f(x_0) - \hat{f}(x_0)\}^2\right] + E(\varepsilon)^2 + 2E\left[\{f(x_0) - \hat{f}(x_0)\}\varepsilon\right]$$

as ε is an Independent Random Number then

$$2E\left[\{f(x_0) - \hat{f}(x_0)\}\varepsilon\right]$$

$$= 2E\left[\{f(x_0) - \hat{f}(x_0)\}\right]\underbrace{E[\varepsilon]}_{0} \Rightarrow 0$$

So $E\left[\{f(x_0) - \hat{f}(x_0)\}^2\right] + E[\varepsilon]^2 + 0$

as we know that $E[\varepsilon]^2 = Var(\varepsilon)$

$\therefore E\left[\{f(x_0) - \hat{f}(x_0)\}^2\right] + Var(\varepsilon) \rightarrow (2)$

= Adding and Subtracting $E(\hat{f}(x_0))$

$\Rightarrow E\left[\{\underbrace{f(x_0) - E(\hat{f}(x_0))}_{a} + \underbrace{E(\hat{f}(x_0)) - \hat{f}(x_0)\}^2}_{b}\right] + Var(\varepsilon) \quad - (3)$

$\Rightarrow E\left[\{\hat{f}(x_0) - E(\hat{f}(x_0))\}^2\right] + E\left[\{E(\hat{f}(x_0)) - \hat{f}(x_0)\}^2\right]$

$+ 2E\left[\{f(x_0) - E(\hat{f}(x_0))\}\{E(\hat{f}(x_0)) - \hat{f}(x_0)\}\right]$

$+ Var(\varepsilon)$

$\Rightarrow E\left[\{\hat{f}(x_0) - E(\hat{f}(x_0))\}^2\right] + E\left[\{E(\hat{f}(x_0)) - \hat{f}(x_0)\}^2\right]$

$+ 2\underbrace{E\left[f(x_0).E(\hat{f}(x_0)) - f(x_0)\hat{f}(x_0) - E^2(\hat{f}(x_0)) + E(\hat{f}(x_0)).\hat{f}(x_0)\right]}_{0}$

$+ Var(\varepsilon)$

$\Rightarrow E\left[\{\hat{f}(x_0) - E(\hat{f}(x_0))\}^2\right] + E\left[\{E(\hat{f}(x_0)) - \hat{f}(x_0)\}^2\right]$

$+ Var(\varepsilon)$

Proof 1

and as we know that from Book

$$\text{Var}(\hat{f}(x_0)) = E\left[\hat{f}(x_0) - E(\hat{f}(x_0))\right]^2$$

and $\text{Bias}(\hat{f}(x_0)) = E\left(f(x_0) - y_0\right)$

and $y_0 = f(x_0) + \underbrace{\varepsilon_0}_{0}$ ($\because$ so $y_0 = f(x_0)$

$$\therefore \underbrace{E\left[\{\hat{f}(x_0) - E(\hat{f}(x_0))\}^2\right]}_{\text{Var}(\hat{f}(x_0))} + \underbrace{\left[\{E(\hat{f}(x_0)) - y_0\}^2\right]}_{\text{Bias}(\hat{f}(x_0))} + \text{Var}(\varepsilon)$$

$$\therefore \Rightarrow \boxed{\text{Var}(\hat{f}(x_0)) + \left[\text{Bias}(\hat{f}(x_0))\right]^2 + \underbrace{\text{Var}(\varepsilon)}_{\text{irreducible error}}}$$

Proof 2

1. Familiarize yourself with the R programming language. Go through 2.3 Lab: Introduction to R (ISLR p. 42–51).

   Solution.
   Done. ■

2. Download the dataset ozone.RData from the course website. (hint: use the load() command). This file contains 3 objects: ozone (the data table), trainset (the row indices for the training set) and testset (the row indices of the test set). Inspect the structure of the objects using ls(), str(), summary(), dim(), length(),range(), colnames(). Identify the column names corresponding to each of the data types mentioned in the introduction. How many observations do you have (in total, in the training set, in the testset)?

   Solution.
   a. Column Names are: "ozone", "radiation", "temperature", "wind"

   b. Data types of the columns:

   - ozone: numeric datatype
   - radiation: integer datatype
   - temperature: integer datatype
   - wind: numeric datatype

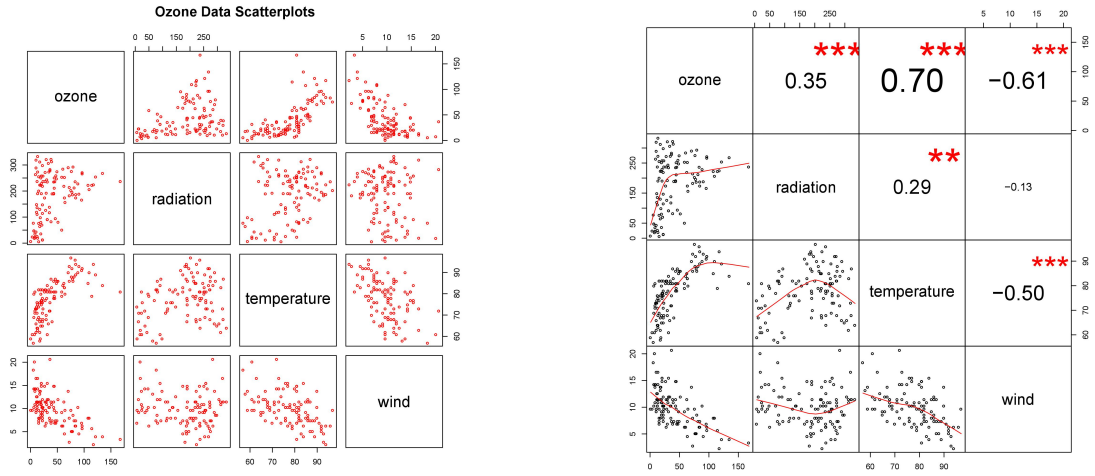
   c. Number of observations: ozone: 111; trainset: 80; testset: 31. ■

3. What is the range of each input variable? What is the mean and standard deviation of each variable?

   Solution.
   Refer Section 3 of the code. The table below shows a summary of the same. ■

	ozone	radiation	wind	temperature
mean	42.1	184.8	77.79	9.939
sd	33.2	91.1	9.52	3.55
range	1-168	7-334	57-97	2.3-20.7

4. Create scatterplots for every pair of features in the dataset. Calculate the Pearson correlation coefficients for each pair of datatypes. In general, what is the range of the Pearson correlation coefficient? What does a correlation coefficient of 0 tell you about the relationship between two variables? What trends do you observe in the data according to the correlation coefficient? Can you see them directly from the plot (visually)?



(a) Scatterplot for every pair of features (b) Correlation trends for every pair of features

Figure 1: Scatterplot and Correlation Trends

Solution. **Figure**1a shows the scatterplot for every pair of features of ozone. **Figure**1b shows the correlation trends for every pair of features for every pair of features of ozone.

a. Range for pearson coefficient is +1 to -1
b. A correlation of zero indicates that the variables do not have a linear association between them.
c. Correlation Trends:

(a) Ozone has a strong positive (0.7) correlation with temperature, moderate correlation with radiation and negatively correlated with wind.

(b) The second variable, radiation, postive correlation with temperature and a small negative correlation with wind.

(c) The third variable, temperature is negatively correlated with wind

■

5. Implement a function rss that computes the Residual Sum of Squares (RSS) between a vector of predicted values and a vector of true values. See 3.6. Lab, section 3.6.7 (ISLR p.119) for how to write R functions.

Solution.
Refer Section 5 of the code. ■

6. Predict the ozone level based on radiation, temperature and wind speed using a linear regression model. Use the training set to train the model and the test set to test the model. Report the RSS as well as the correlation (Pearson) with the true responses. Create a scatterplot for the predicted and true values of ozone for the test set.

Solution.
a. The RSS value of lm model: 8208.509. Pearson Coeff for lm model is: 0.8268958.

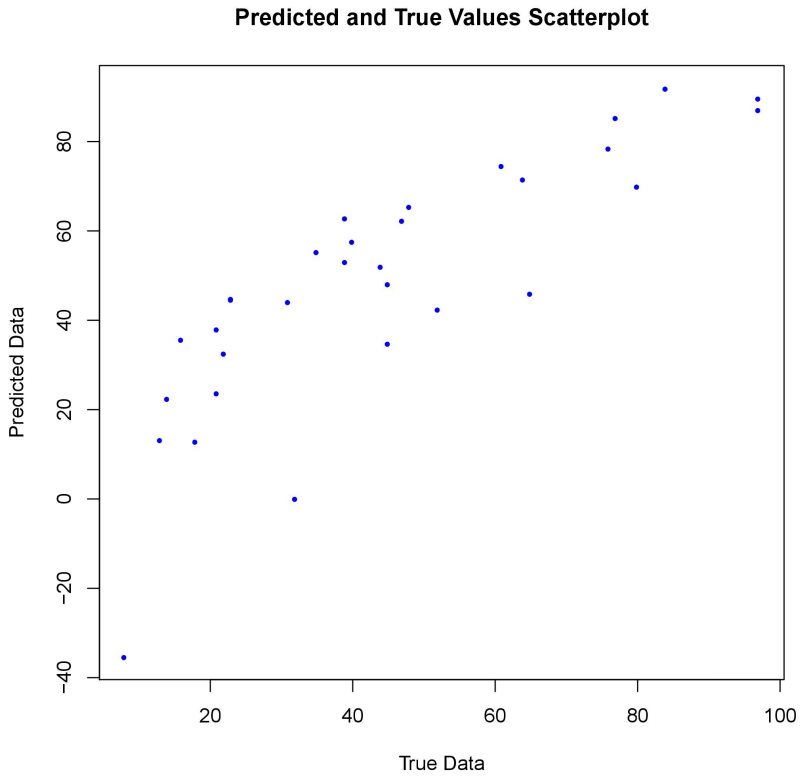


Figure 2: Scatter Plot between predicted and trues values of ozone using linear regression model

Figure2 shows the scatter Plot between predicted and trues values of ozone using linear regression model. ■

7. Perform k nearest neighbor (kNN) regression to predict the ozone level from the other features. Use k = 1, 2, . . . , 30. Plot the RSS for training and test set respectively for every k. On which side of the graph do you have the most complex models? Argue with the bias-variance tradeoff. Which value of k would you choose for this data? In general, does the kNN method make any assumptions on the underlying data distribution?

Solution.
Refer Section 5 of the code. **Figure**3 shows the Plot the RSS for training and test set respectively for every k using k-nearest neighbour algorithm.

Model complexity of a kNN model increases as the number of neighbours increase. Consequently right side of the graph would have the most complex model. Consider the most simple model, k=1, this model would fit perfectly to the nearest point hence

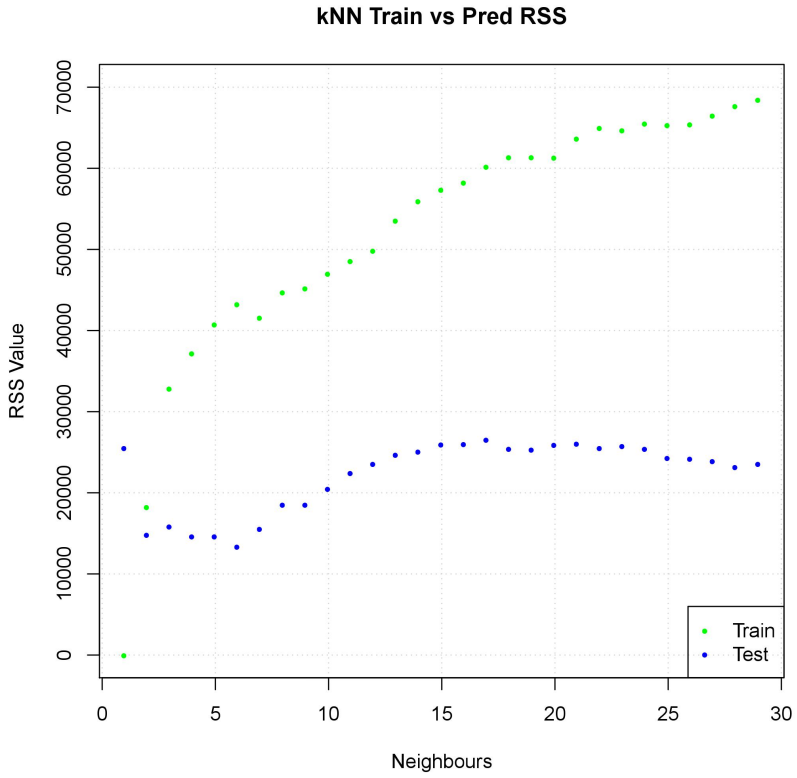


Figure 3: RSS plot for training and test set respectively for every k

a very low train error or low bias. This would not generalize to the unseen data,thus might lead to high variance/ error with respect to a new point from the test set.

In general, kNN method does not make any assumptions about the underlying data distribution.

But it follows a philosophy that similar data points are close to each other, in other words, neighbours. Further, in kNN models, the features are in metric space and hence have a notion of distance between different data points. For example, $\mathbb{R}^2$ space is commonly chosen with euclidean distance as a measure of distance. ■

8. Compare the RSS and correlation of the linear model and your chosen nearest neighbor model. Which one would you prefer over the other for this example? Consider model complexity (degrees of freedom), model assumptions and prediction quality.

Solution.
The RSS and correlation of the linear model and kNN model are as below:

- The RSS value of best kNN model: 13417
  Pearson Coeff for best kNN model is: 0.6052382

- The RSS value of lm model: 8208.509
  Pearson Coeff for lm model is: 0.8268958

- **Model Complexity:** Linear regression model is simple linear relation between input features while the best k-NN model a complex relationship of considering multiple neighbours, for this problem, 6 nearest neighbours. Hence, lm model is simpler relative to k-NN approach.
- **Model Assumptions:** Linear Regression model assumes a linear relation between input and the target variables using a finite set of parameters (in this case 4). k-NN model does not make any assumptions on the underlying data distribution. For the given ozone data it the pearson coefficients indicate a linear relationship between the variables, hence assuming a linear relationship would be benificial.
- **Prediction quality:** The test error i.e., RSS value for the lm model prediction is lower than the k-NN model. Hence, the linear model has better prediction quality.

Clearly it is evident that the linear regression model outperforms the best kNN model by a good margin. Hence, for this example, Linear Regression model is the best choice. ■