Chapter 2

Statistical Learning

STAT303-2

Northwestern | WEINBERG COLLEGE OF ARTS & SCIENCES

Check the hand-written notes

Recap on Datasets

Remember that in a dataset:

- Each row is a data point/instance/observation, X
- Each column is a feature/predictor

$X_1, X_2, \ldots, X_p$

One of the columns in the dataset (or a new column overall) will be predicted using the X columns, **the predictors**. This is the **response or the dependent/target variable**: **Y**

Check the hand-written notes

Statistical learning

lin.

If we observe a quantitative response Y, and p different predictors $X_1, X_2, \ldots, X_p$, we assume that there is some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$, which can be written in the general form:

$$Y = \underbrace{f(X)}_{\text{A function}} + \underbrace{\epsilon}_{\text{Noise: Generated by a random variable}}$$

For example, Y may be the price of a car, and X may consist of features such as mileage, age, model, etc.

The real f is unknown and cannot be found out; it can be estimated as $\hat{f}$, using the existing data.

- The predictors, X
- The existing target values, Y

But why do we want to estimate f ?

Check the hand-written notes

Statistical learning: Purpose of estimating f

There are two main reasons for estimating *f*:

a. **Prediction**

We are interested in predicting the response (or the dependent variable).

$$\hat{Y} = \hat{f}(X),$$

where $\hat{f}$ is the estimate of *f*, and $\hat{Y}$ is the prediction for Y.

For example, we wish to predict the price of a car based on its features.

b. **Inference**

We are interested in identifying the relationship between the predictor(s) and the response.

For example, we wish to identify:

1. Which car features are associated with its price *(or have a statistically significant relationship with its price)*?
2. How much increase in mileage is associated with a given decrease in car price?

Statistical learning: Techniques for estimating *f*

It can be shown that:

$$E\left(Y-\hat{Y}\right)^2 = E\left[f(X)+\in-\hat{f}(X)\right]^2$$

$$=> \;\; E\left(Y-\hat{Y}\right)^2 = \underbrace{\left[f(X)-\hat{f}(X)\right]^2}_{\text{Reducible error}} + \underbrace{Var(\in)}_{\text{Irreducible error}}$$

In both Data Science II & III, we'll learn techniques for estimating *f* with the aim of minimizing the reducible error

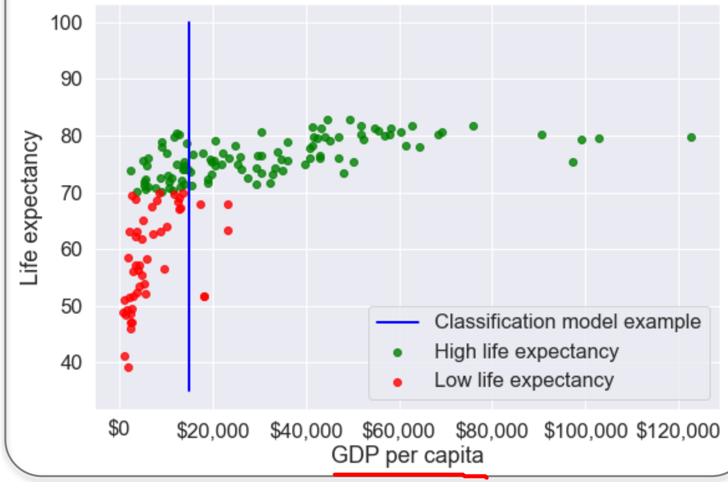
Regression vs classification

Regression:

- The response (Y) is a continuous variable. For example, predicting life expectancy of a country (Y) based on its GDP per capita. (X)

Classification:

- The response (Y) is a categorical variable. For example, classifying a country as having low (0) or high (1) life expectancy based on its GDP per capita.



Regression vs classification

Regression:

- The response (Y) is a continuous variable. For example, predicting life expectancy of a country based on its GDP per capita.
- Typically, the regression model directly predicts the continuous response.

Classification:

- The response (Y) is a categorical variable. (a class) For example, classifying a country as having low (0) or high (1) life expectancy based on its GDP per capita.
- Typically, the classification model predicts a probability of response *(or probability of the observation belonging to one of the classes)*. The class is then predicated based on a user-defined threshold probability.

Assessing model accuracy: Regression

There are several metrics that can be used to assess the prediction accuracy of a regression model. Below are a couple of popular ones:

1. RMSE (Root mean squared error)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{f}(X^i))^2}$$

2. MAE (Mean absolute error)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|Y_i - \hat{f}(X^i)\right|$$

Note that the superscript 'i' in the above formulae denotes the i^{th} observation, and N denotes the total number of observations (or rows) in the data.

Assessing model accuracy: Regression

Which one to choose for a given problem: RMSE or MAE? → Depends on how you want to penalize the errors

1. RMSE (Root mean squared error)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{f}(X^i))^2}$$

- Each error is squared before adding to the total error sum
- Large errors are penalized more.
- Errors between 0 and 1 are penalized less.

2. MAE (Mean absolute error)

$$MAE = \frac{1}{N}\sum_{i=1}^{n}\left|Y_i - \hat{f}(X^i)\right|$$

The error from each observation is equally penalized.

Assessing model accuracy: Classification

- In case of binary classification, the most basic metric is accuracy:
- For more advanced analysis, a confusion matrix can be generated as shown below.

$$Accuracy = \frac{\#\ Correctly\ predicted\ instances}{\#\ All\ instances} \times 100$$

Confusion matrix	Predicted: 0	Predicted: 1
Actual: 0	TN	FP
Actual: 1	FN	TP

- Several metrics for quantifying model accuracy are based on the confusion matrix.

- A popular metric for quantifying the overall classification accuracy is the classification error rate:

$$Classification\ error\ rate = \frac{FN + FP}{TN + TP + FN + FP} = 1 - \text{Accuracy}$$

Assessing model accuracy: Classification

Confusion matrix	Predicted: 0	Predicted: 1
Actual: 0	TN	FP
Actual: 1	FN	TP

- Sometimes, the overall classification error rate may not suffice in assessing the utility of the model and the risks associated in case of misclassification

- The metrics below shed light on the accuracy of the model in different cases:

$$False\ positive\ rate\ (FPR) = \frac{FP}{FP + TN}$$

$$False\ negative\ rate\ (FNR) = \frac{FN}{FN + TP}$$

$$Precision = \frac{TP}{FP + TP}$$

$$Recall\ or\ TPR\ = 1 - FNR = \frac{TP}{FN + TP}$$

Assessing model accuracy: Classification

Confusion matrix	Predicted: 0	Predicted: 1
Actual: 0	TN	FP
Actual: 1	FN	TP

- All these metrics are ratios – all between 0 and 1.
- FPR and FNR should be low – closer to 0
- Precision and Recall should be high – closer to 1
    - A model that predicts most observations as 1: Low precision, high recall
    - A conservative model that predicts very few observations as 1: High precision, low recall
    - Ideal scenario: High precision, high recall – low number of FP and FN

$$False\ positive\ rate\ (FPR) = \frac{FP}{FP + TN}$$

$$False\ negative\ rate\ (FNR) = \frac{FN}{FN + TP}$$

$$Precision = \frac{TP}{FP + TP}$$

$$Recall\ or\ TPR\ = 1 - FNR = \frac{TP}{FN + TP}$$

Assessing model accuracy: Classification

- All the metrics can be obtained for a given classification model. However, while developing the model, some metrics may be more important than others.
- Suppose the classification problem is to predict if a person has diabetes $(y = 1)$ or does not have diabetes $(y=0)$ based on their symptoms.

Which is the most important metric to assess the accuracy of this classification model?

- It may be worse to classify a person having diabetes $(y = 1)$ as not having diabetes $(y = 0)$, as opposed to the case where the person not having diabetes $(y = 0)$ is classified as having diabetes $(y = 1)$.
- Thus, in this particular case, reducing FNR the most important metric.
- However, FPR shouldn't be too high, and precision shouldn't be too low, or the model will cease to be useful.
- Thus, in this case, one should try to develop a model with a low FNR, but also having a reasonable FPR and precision.

Assessing model accuracy: Classification

Some other popularly used metrics to assess a classification model accuracy that we'll see later in detail in the course are:

- Precision - recall
- ROC-AUC (Receiver operating characteristic – Area under the curve)

Reference

Source for slides: https://www.statlearning.com/resources-second-edition