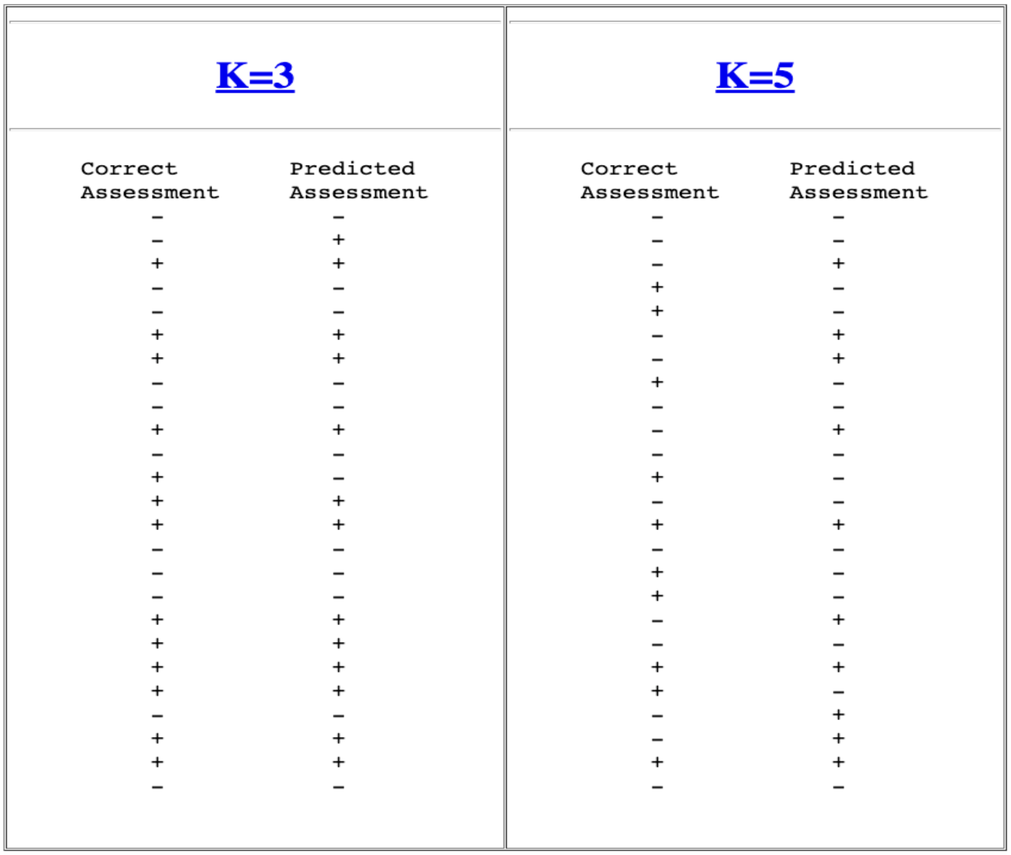
KNN + Confusion Matrix

CS550 Homework
Shoumya Singh
ID-19566

29. KNN + Confusion Matrix

Evaluation Phase

- Objective
  - Finding the K value representing the best model.
- How? -- This is the homework you need to do.
  - Using Pick an Evaluation Metric: Confusion Matrix
  - For example, for credit card assessment
    - There are two classes in this example, "+" (credit approval) and "-" (credit denial).



K=3		K=5	
Correct Assessment	Predicted Assessment	Correct Assessment	Predicted Assessment
-	-	-	-
-	+	-	-
+	+	-	+
-	-	+	-
-	-	+	-
+	+	-	+
+	+	-	+
-	-	+	-
-	-	-	-
+	+	-	+
-	-	-	-
+	-	+	-
+	+	-	-
+	+	+	+
-	-	-	-
-	-	+	-
-	-	+	-
+	+	-	+
+	+	-	-
+	+	+	+
+	+	+	-
-	-	-	+
+	+	-	+
+	+	+	+
-	-	-	-

- If the objective is to determine the "+" class, please fill this table

K=	TP	FN	FP	TN	Precision	Accuracy	Recall	F1 score
3								
5								

- Which K value represents the better model? Please explain your assessment.

Solution:

When **N = 25, K= 3**

N = 25, K= 3	**Predicted +**	**Predicted -**
Correct +	12 TP	1 FP
Correct -	1 FN	11 TN

$$\textbf{Accuracy} = \frac{\mathbf{TP + TN}}{\mathbf{TP + TN + FP + FN}}$$

$$= \frac{\mathbf{12 + 11}}{\mathbf{12 + 11 + 1 + 1}} = \frac{\mathbf{23}}{\mathbf{25}} = \mathbf{0.92}$$

Precision = **Positive Predictive Value (PPV)** = $\dfrac{\mathbf{TP}}{\mathbf{TP + FP}}$

$$= \frac{\mathbf{12}}{\mathbf{12 + 1}} = \frac{\mathbf{12}}{\mathbf{13}} = \mathbf{0.92}$$

Recall = Sensitivity

= Out of all the positive data points, how many have been truly identified as positive

= Hit Rate

= True Positive Rate (TPR) $= \frac{TP}{\underline{P}} = \frac{TP}{TP + FN}$

$$= \frac{12}{12 + 1} = 0.92$$

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = 2 * \frac{\text{presicion} * \text{recall}}{\text{presicion} + \text{recall}} = \frac{TP}{TP + \frac{FN + FP}{2}}$$

$$\text{F1 Score} = \frac{12}{12 + \frac{1 + 1}{2}} = \frac{12}{13} = 0.92$$

When **N = 25, K= 5**

N = 25, K= 5	**Predicted +**	**Predicted -**
Correct +	3 TP	7 FP
Correct -	7 FN	8 TN

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$= \frac{3 + 8}{3 + 8 + 7 + 7} = \frac{11}{25} = 0.44$$

Precision = **Positive Predictive Value** (PPV) = $\frac{TP}{TP + FP}$

$$= \frac{3}{3 + 7} = \frac{3}{10} = 0.3$$

Recall = **Sensitivity**

= Out of all the positive data points, how many have been truly identified as positive

= **Hit Rate**

= True Positive Rate (TPR) = $\frac{TP}{\underline{P}} = \frac{TP}{TP + FN}$

$$= \frac{3}{3 + 7} = 0.3$$

$$\boxed{\text{F1 Score} = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = 2 * \frac{\text{presicion} * \text{recall}}{\text{presicion} + \text{recall}} = \frac{TP}{TP + \frac{FN + FP}{2}}}$$

$$\text{F1 Score} = \frac{3}{3 + \frac{7 + 7}{2}} = \frac{3}{10} = 0.3$$

K=	TP	FN	FP	TN	Precision	Accuracy	Recall	F1 score
3	12	1	1	11	0.92	0.92	0.92	0.92
5	3	7	7	8	0.3	0.44	0.3	0.3

K = 3 is a better model, because the classifier will only get a high F1 score if both recall, and precision are high. In this case the F1 score of K = 3 is high.