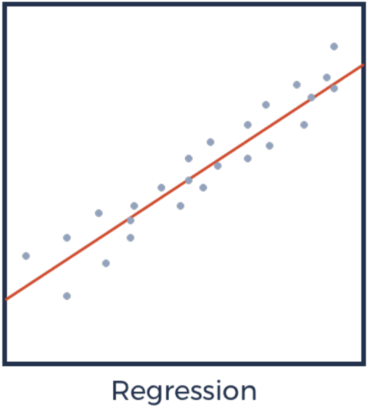
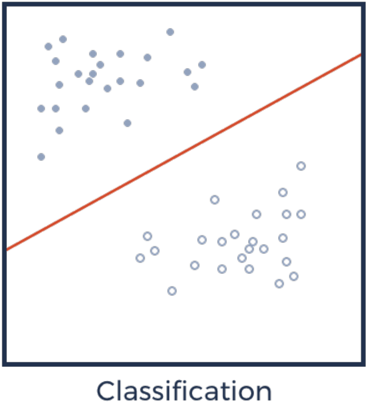
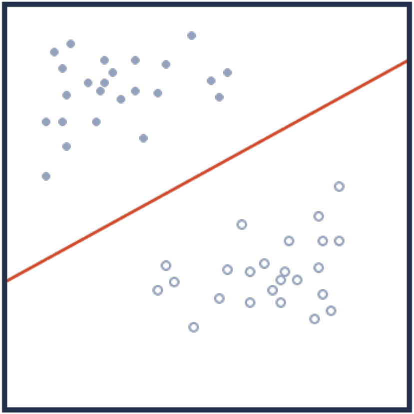
Supervised Learning

Repeating the mistakes of the past







Classification

Confusion Matrix

c_{ij} is the count of data from class *i* that was labeled as class *j*

$$
\begin{array}{c}
\text{Predicted} \\
\mathbf{C} = \begin{array}{c c}
 & \begin{array}{ccccc} 1 & 2 & 3 & \dots & n \end{array} \\
\begin{array}{c} 1 \\ 2 \\ \vdots \\ n \end{array} &
\begin{bmatrix}
c_{11} & c_{12} & c_{13} & \dots & c_{1n} \\
c_{21} & c_{22} & c_{23} & \dots & c_{2n} \\
\vdots & \ddots & & & \vdots \\
c_{n1} & c_{n2} & c_{n3} & \dots & c_{nn}
\end{bmatrix}
\end{array} \text{Actual}
\end{array}
$$

Accuracy

$$\mathbf{C} = \begin{array}{cc} \text{Predicted} & \\ \begin{bmatrix} c_{11} & c_{12} & c_{13} & \dots & c_{1n} \\ c_{21} & c_{22} & c_{23} & \dots & c_{2n} \\ \vdots & \ddots & & & \vdots \\ c_{n1} & c_{n2} & c_{n3} & \dots & c_{nn} \end{bmatrix} & \text{Actual} \end{array}$$

$$\text{Accuracy} = \frac{\sum_i c_{ii}}{\sum_i \sum_j c_{ij}}$$

$$\mathbf{C} = \begin{array}{cc} \text{Predicted} & \\ \begin{bmatrix} c_{11} & c_{12} & c_{13} & \dots & c_{1n} \\ c_{21} & c_{22} & c_{23} & \dots & c_{2n} \\ \vdots & \ddots & & & \vdots \\ c_{n1} & c_{n2} & c_{n3} & \dots & c_{nn} \end{bmatrix} & \text{Actual} \end{array}$$

$$\text{Accuracy} = \frac{\mathbf{tr}(\mathbf{C})}{\sum_i \sum_j c_{ij}}$$

True Positive, True Negative

$$\mathbf{C} = \begin{array}{c} \text{Predicted} \\ \begin{bmatrix} c_{11} & c_{12} & c_{13} & \dots & c_{1n} \\ c_{21} & c_{22} & c_{23} & \dots & c_{2n} \\ \vdots & \ddots & & & \vdots \\ c_{n1} & c_{n2} & c_{n3} & \dots & c_{nn} \end{bmatrix} \end{array} \text{Actual}$$

$$TP_i = c_{ii}$$

$$TN_i = \sum_{j \neq i, k \neq i} c_{jk}$$

True Positive, True Negative

$$\mathbf{C} = \begin{array}{cc} \text{Predicted} & \\ \begin{bmatrix} c_{11} & c_{12} & c_{13} & \cdots & c_{1n} \\ c_{21} & c_{22} & c_{23} & \cdots & c_{2n} \\ \vdots & \ddots & & & \vdots \\ c_{n1} & c_{n2} & c_{n3} & \cdots & c_{nn} \end{bmatrix} & \text{Actual} \end{array}$$

$$TP_i = c_{ii}$$

$$TN_i = \sum_{j \neq i, k \neq i} c_{jk}$$

False Positive, False Negative

$$\begin{array}{cc} & \text{Predicted} \\ \mathbf{C} = & \begin{bmatrix} c_{11} & c_{12} & c_{13} & \dots & c_{1n} \\ c_{21} & c_{22} & c_{23} & \dots & c_{2n} \\ \vdots & \ddots & & & \vdots \\ c_{n1} & c_{n2} & c_{n3} & \dots & c_{nn} \end{bmatrix} & \text{Actual} \end{array}$$

$$FP_i = \sum_{j \neq i} c_{ji}$$

$$FN_i = \sum_{j \neq i} c_{ij}$$

False Positive, False Negative

$$\begin{array}{ccc} & \text{Predicted} & \\ \mathbf{C} = & \begin{bmatrix} c_{11} & c_{12} & c_{13} & \cdots & c_{1n} \\ c_{21} & c_{22} & c_{23} & \cdots & c_{2n} \\ \vdots & \ddots & & & \vdots \\ c_{n1} & c_{n2} & c_{n3} & \cdots & c_{nn} \end{bmatrix} & \text{Actual} \end{array}$$

$$FP_i = \sum_{j \neq i} c_{ji}$$

$$FN_i = \sum_{j \neq i} c_{ij}$$

Summary

$$\begin{array}{cc} & \text{Predicted} \\ \mathbf{C} = & \begin{bmatrix} c_{11} & c_{12} & c_{13} & \cdots & c_{1n} \\ c_{21} & c_{22} & c_{23} & \cdots & c_{2n} \\ \vdots & \ddots & & & \vdots \\ c_{n1} & c_{n2} & c_{n3} & \cdots & c_{nn} \end{bmatrix} & \text{Actual} \end{array}$$

$$TP_i = c_{ii}, \qquad FP_i = \sum_{j \neq i} c_{ji},$$

$$TN_i = \sum_{j \neq i, k \neq i} c_{jk}, \qquad FN_i = \sum_{j \neq i} c_{ij}$$

Precision

$$TP_i = c_{ii}, \qquad FP_i = \sum_{j \neq i} c_{ji},$$

$$TN_i = \sum_{j \neq i, k \neq i} c_{jk}, \qquad FN_i = \sum_{j \neq i} c_{ij}$$

$$Pr_i = \frac{TP_i}{TP_i + FP_i}$$

Recall

$$TP_i = c_{ii}, \qquad FP_i = \sum_{j \neq i} c_{ji},$$

$$TN_i = \sum_{j \neq i, k \neq i} c_{jk}, \qquad FN_i = \sum_{j \neq i} c_{ij}$$

$$Re_i = \frac{TP_i}{TP_i + FN_i}$$

F-1 Score

$$Pr_i = \frac{TP_i}{TP_i + FP_i} \qquad Re_i = \frac{TP_i}{TP_i + FN_i}$$

$$F1_i = \frac{2 \times Pr_i \times Re_i}{Pr_i + Re_i}$$

Class Metrics Summary

$$TP_i = c_{ii}, \qquad FP_i = \sum_{j \neq i} c_{ji},$$

$$TN_i = \sum_{j \neq i, k \neq i} c_{jk}, \qquad FN_i = \sum_{j \neq i} c_{ij}$$

$$Pr_i = \frac{TP_i}{TP_i + FP_i}, \qquad Re_i = \frac{TP_i}{TP_i + FN_i}$$

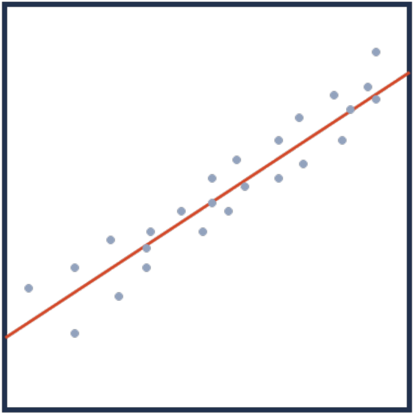
$$F1_i = \frac{2 \times Pr_i \times Re_i}{Pr_i + Re_i}$$

Summary Metrics

$$Pr_{micro} = \frac{\sum_i TP_i}{\sum_i (TP_i + FP_i)}$$

$$Pr_{macro} = \frac{\sum_i Pr_i}{n}$$

$$Pr_{weighted} = \sum_i \left(\frac{c_i}{\sum_j c_j} Pr_i\right).$$



Regression

Mean Squared Error

$$\text{MSE} = \frac{\sum_i (y_i - f_i)^2}{n}$$

$$\text{RMSE} = \sqrt{\frac{\sum_i (y_i - f_i)^2}{n}}$$

R^2

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \hat{y})^2},$$

Methods

Demo

Chet Haase
@chethaase

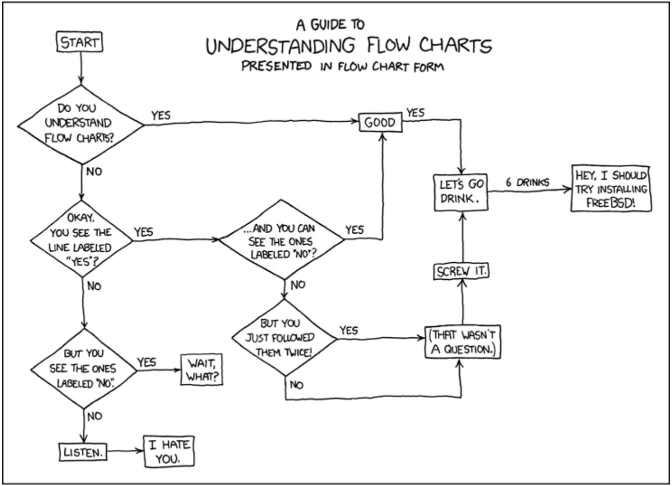
A Machine Learning algorithm walks into a bar.
The bartender asks, "What'll you have?"
The algorithm says, "What's everyone else having?"

7:24 AM - Nov 1, 2017

8,486 4,342 people are talking about this

k Nearest Neighbors Considerations

- kNN delays the computational effort until the inference/prediction stage
- k is a hyperparameter that can affect the model's performance
- The way we calculate "nearest points" can significantly affect how our model performs. We can use an ℓ_p-norm for our distance metric and ℓ_1, and ℓ_2 are usually effective.



https://xkcd.com/518/

Impurity Metrics

$$\text{Gini}_m = \sum_i p_{mi}(1 - p_{mi})$$

$$\text{Cross-entropy}_m = -\sum_i p_{mi} \log p_{mi}$$

Decision Trees

- The splitting criterion or measure for impurity is a hyperparameter
- The stopping criterion for training is a critical hyperparameter to prevent overfitting of the data

Random Forests

An ensemble of decision tree models.

Random Forests Considerations

Same as decision trees plus,

- the number of decision trees used to create the random forest model is a hyperparameter to consider.

Questions