Welcome to week 5!

Topics to cover

- Supervised learning
  - Forms of supervised learning algorithm
  - Example of a supervised learning
- Model complexity
  - Concept of model complexity
  - Model complexity and Occam’s razor
  - Structural risk minimisation
- From week 6!
  - Model evaluation metrics

Forms of Supervised Learning

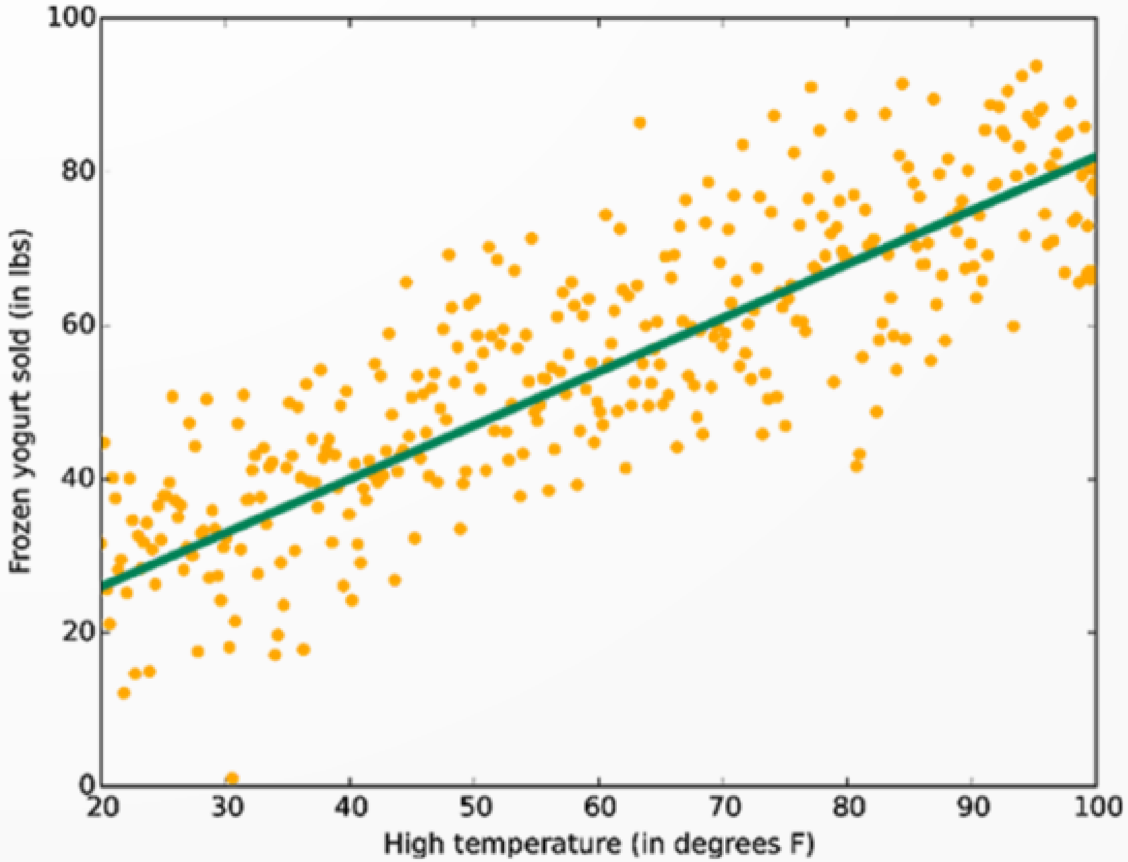
- Supervised learning is
  - Learning the mapping function that maps the input variable x to the output variable y.
  - estimating a function from labelled training data i.e., $y_i = h(x_i)$.
- The data used to train the algorithm is already labelled with correct answers i.e., $\{x_i, y_i\}, i = 1, ..., n$
- The majority of practical machine learning applications, use supervised learning.
- Benefits of supervised learning:
  - Instead of finding patterns based on similarity only, we can learn a direct mapping or function between feature vector x_i and the output (target or label) y_i

Forms of Supervised Learning...

- Supervised learning can appear in many forms:
  - Regression problem
    - Linear Regression (linear model)
    - Logistic Regression (linear model)
  - Classification problem
    - Support Vector Machines (both linear and nonlinear)
    - Decision Trees (nonlinear)
    - Random Forest (nonlinear)
    - Neural Networks: Perceptron and Multi-layer Perceptron (nonlinear)
  - Ranking problem

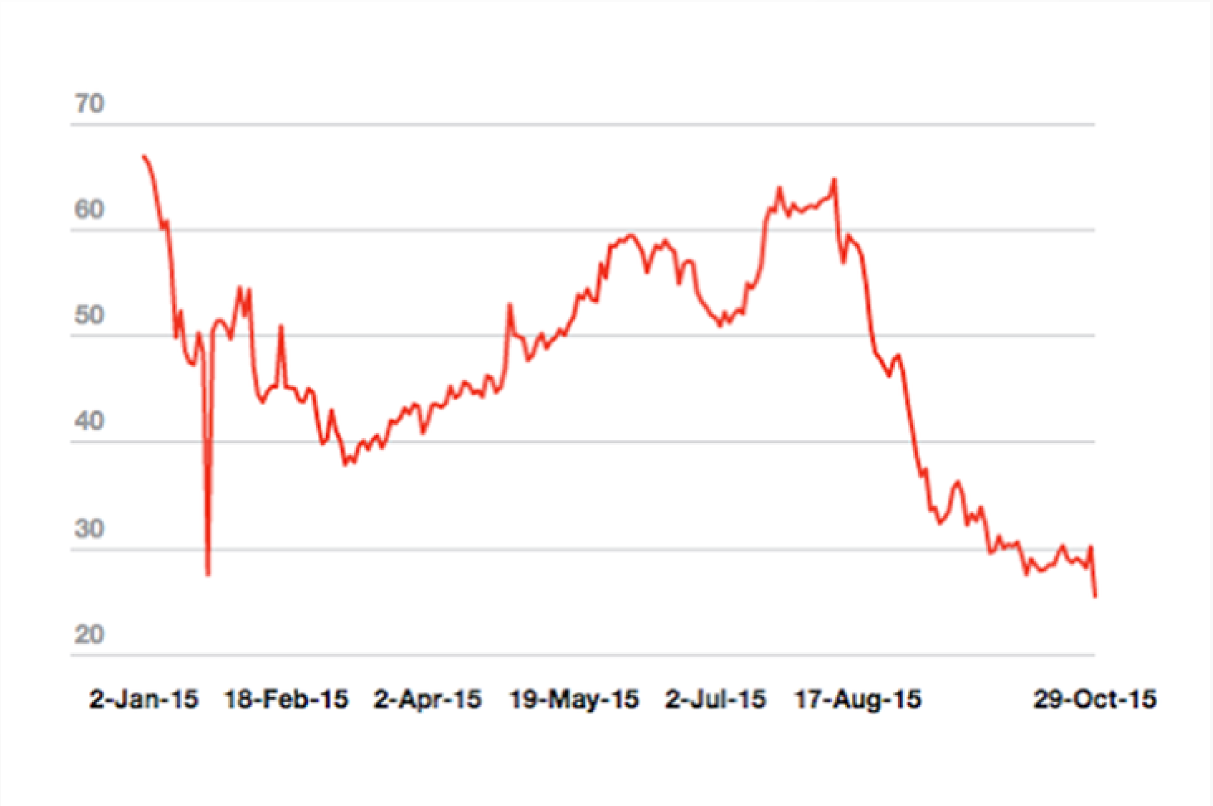
Forms of Supervised Learning...

- Example-1:
  - The following figure illustrates a regression problem about the sale of Yogurt with seasonal temperature.
  - Let's estimating the relationships among the feature variables (e.g. the sale of frozen yogurt and its temperature).



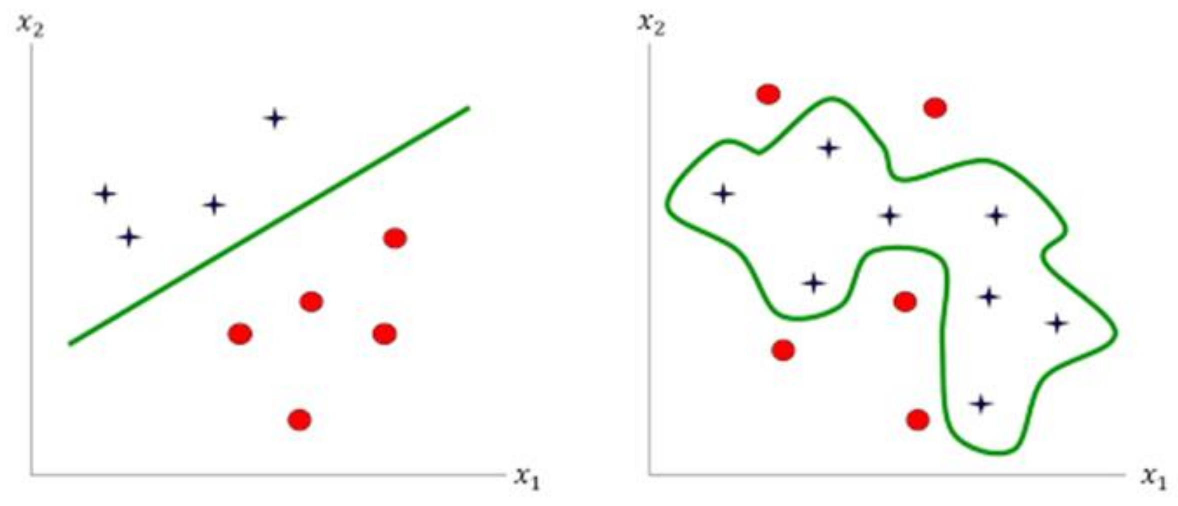
Forms of Supervised Learning...

- Example-2: *Regression*
  - We look to find relationships among feature variables.
  - The figure illustrates sample data for GoPro stock price against date.
  - Imagine the amount of money you can earn by intelligently predicting prices in the stock market!



Forms of Supervised Learning...

- Example-2:
  - The following figure illustrates a classification problem for classifying two types of data.
  - As you can see, sometimes we can successfully find a linear boundary and sometimes we have to search for a more complex boundary.



Linear decision boundary *Nonlinear decision boundary*

How Supervised Learning Algorithm Works

- Consider a supervised learning algorithm with *n* training data $\{x_i, y_i\}, i = 1, ..., n$
  - The learning algorithm seeks a function on $h : X \rightarrow Y$

    where X is the input space and Y is the output space.
- The function h is:
  - an element of some space of possible functions H, usually called the hypothesis space.
- Start with a hypothesis function that we think is similar to the true function behind the data.
  - End up with a function as accurate as possible to the main unknown function.

How Supervised Learning Algorithm Works...

- How can we measure the quality of function h?
- How can we understand how accurate h can map X to the target Y?
- To answer this question, we need to introduce a new function called the **loss function**.
  - A function h is applied to a training instance x_i and it gives the output $h(x_i)$, so that $\hat{y}_i = h(x_i)$
    - Since we are dealing with a supervised problem we know that the true output is y_i
    - In order to measure how well a function h fits the training data, a loss function $L(y_i, \hat{y}_i)$ between y_i and $\hat{y}_i$ is defined.

How Supervised Learning Algorithm Works...

- Some familiar loss functions:
- Square loss:

  $$L(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$$ (useful for regression)

- Absolute loss:

  $$L(y_i, \hat{y}_i) = |y_i - \hat{y}_i|$$ (useful for regression)

- 0-1 loss: $L(y_i, \hat{y}_i) = 1(y_i, \hat{y}_i)$ which is equal to 0 if $y_i = \hat{y}_i$ and 1 otherwise (useful for classification)

- Other loss functions for classification problem:
  e.g. Logistic loss, Hinge loss

How Supervised Learning Algorithm Works…

- The loss function is used to compute the error between the actual result of y_i and what we calculated $\hat{y}_i$
- Similar to the loss function, we can define a factor called empirical risk:
  - average of the loss function $$\frac{1}{n}\sum_{i=1}^{n} L(y_i, h(x_i))$$

How Supervised Learning Algorithm Works…

- Among all functions in hypothesis space, that is, $h \in H$, we select the function h, which minimizes the empirical risk.
  - But how can achieve this? Minimize the risk of loss!!!
  - In other words, we select a function h that achieves minimum risk:

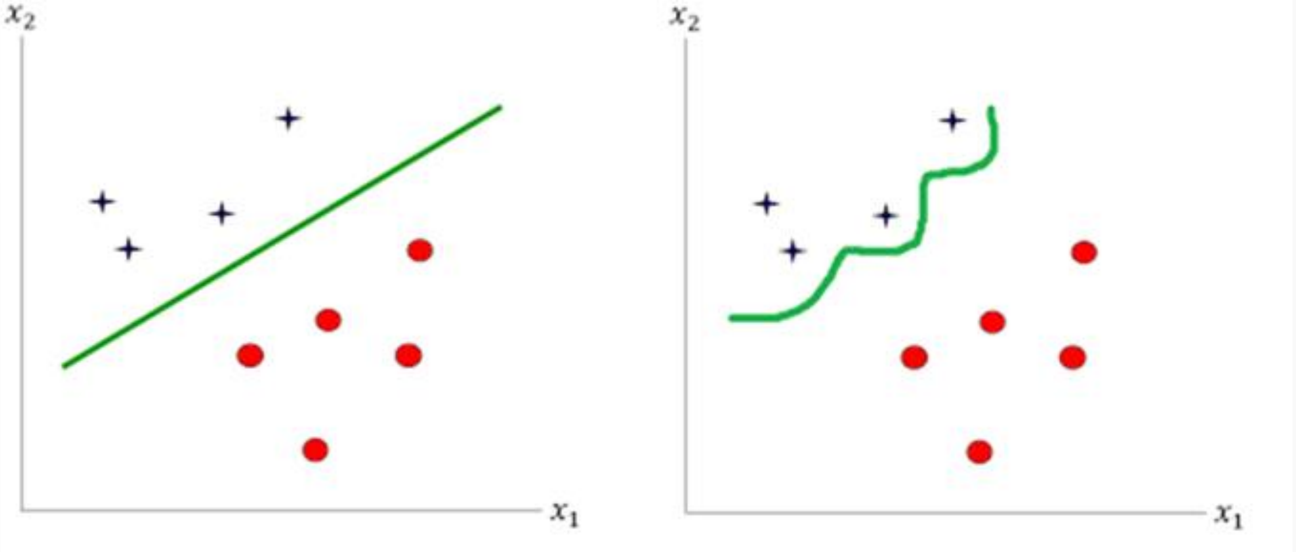
$$\min_{h \in H} \frac{1}{n} \sum_{i=1}^{n} L(y_i, h(x_i))$$

Model Complexity

Concept
Occam’s Razor
Structural Risk Minimisation

Concept of Model Complexity

- How complex should a machine learning model be?
- What are the costs when a complex model is used? When is it necessary to use a complex model?
- In this classification problem, which boundary line seems more appropriate?



Linear decision boundary　　*Nonlinear decision boundary*

Concept of Model Complexity...

- We may not always be able to visualise the training data in high dimensions.
- So, we may not know whether the regression problem is linear or non-linear.
- What should be the right complexity of the model that we use to fit the given data?
  - Effects of selecting different models in terms of complexity.

Concept of Model Complexity…

- The effects of selecting different models in terms of complexity:
  - If we choose higher complexity than necessary, we would be over-fitting the data.
  - If we choose lower complexity than necessary, we would be under-fitting the data.
  - It is important to get the right fit for good generalisation.

It is prediction on unseen data, that is, the data, which are not part of our training set

Model Complexity & Occam's Razor

- Occam's Razor (a famous problem-solving principle) is used as a heuristic guide in the development of theoretical models.
  - "**All other things being equal, the simplest solution is the best**"
- It also addresses the problem of which hypothesis to choose if there are multiple hypothesis with similar fit.

Structural risk minimisation

- *Based on Occam's razor* and its simplistic principle, we *define another risk value* which is called **Structural Risk**.
- Structural risk minimisation seeks to prevent over-fitting by incorporating a penalty on the model complexity that prefers simpler functions over more complex ones.
- The general idea is to minimise both Structural Risk and Empirical Risk

$$R_{str}(h) = R_{emp}(h) + \lambda C(h)$$

Where $C(h)$ is the complexity of hypothesis function h and λ is a penalty parameter.

Model Evaluation Metrics

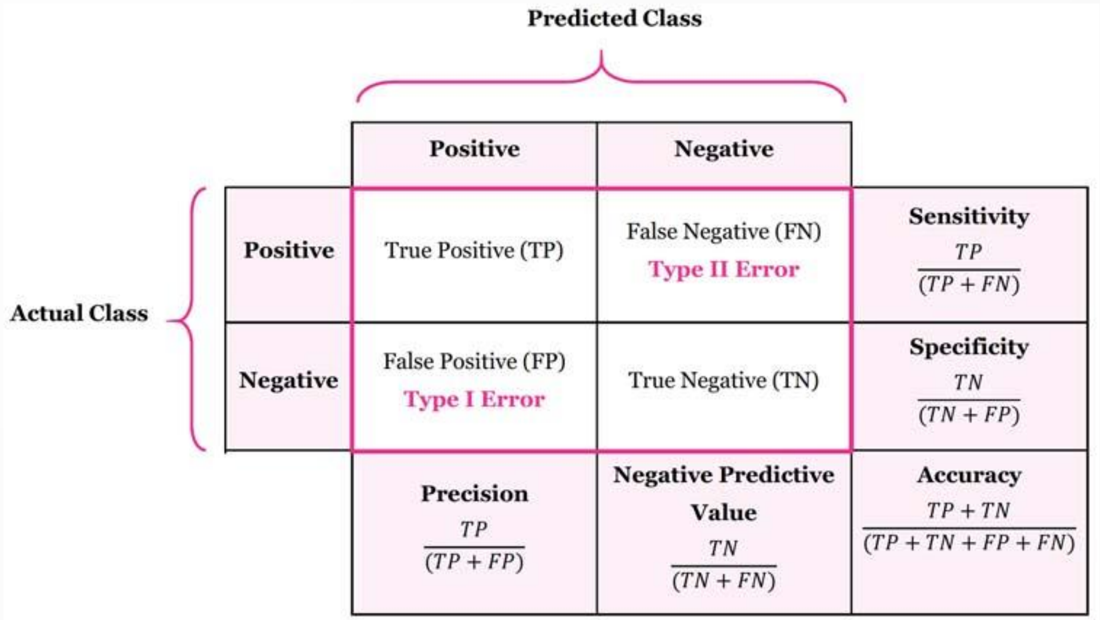
Classification Metrics
Regression Metrics

Classification Metrics

- The metrics that you choose to evaluate your machine learning model is very important
  - The choice of metrics influences how the performance is measured and compared
- There are a myriad of metrics that can be used to evaluate predictions for classification problems
  - Confusion Matrix
  - ROC Curve
  - F-1 Measure

Classification Metrics…
Confusion Matrix

- A confusion matrix is a summary of prediction results on a classification problem
  - The number of correct and incorrect predictions are summarized with count values and divided down by each class.
  - Confusion matrices are a way to understand the types of errors made by a model.



Classification Metrics… Confusion Matrix

- Accuracy is not reliable!
  - Use confusion matrix
    - Unbalanced data (i.e. when the numbers of observations in different classes vary greatly)
      - Accuracy may generate confusing result
      - For example, if there were 90 apples and only 10 oranges in the data set, a particular classifier might classify all the observations as apples.
        - Is this wise?

Classification Metrics…
Confusion Matrix

prediction

truth	Cat(a)	Cat(b)	Cat(c)	Cat(d)	Cat(e)	Cat(f)	Cat(g)	Cat(h)	Cat(i)
Cat(a)	90%	1%	0%	0%	0%	0%	0%	0%	0%
Cat(b)	6%	94%	2%	0%	0%	0%	0%	0%	1%
Cat(c)	2%	2%	94%	0%	0%	0%	0%	2%	1%
Cat(d)	0%	0%	0%	92%	0%	1%	0%	0%	0%
Cat(e)	2%	0%	0%	0%	90%	1%	0%	0%	0%
Cat(f)	0%	0%	0%	0%	5%	93%	0%	4%	1%
Cat(g)	0%	2%	0%	1%	0%	0%	91%	0%	0%
Cat(h)	0%	0%	0%	3%	2%	1%	0%	90%	1%
Cat(i)	0%	0%	2%	0%	1%	1%	3%	0%	94%
Cat(j)	0%	1%	1%	2%	0%	0%	5%	2%	2%
Cat(k)	0%	0%	1%	2%	2%	3%	1%	2%	0%

- The higher the proportion of values on the diagonal of the matrix in relation to values off of the diagonal, the better the classifier is (why?).

Classification Metrics…
Confusion Matrix

- Consider the following figure as a confusion matrix for only two classes.
- You could represent the positive class as class 1 and the negative class as class 0.

		actual class	
		positive	negative
predicted class	positive	true positives (TP)	false positives (FP)
	negative	false negatives (FN)	true negatives (TN)

- In this case we define the accuracy as:

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

- But as we have said before, accuracy may not be a useful metric for imbalanced class problems.

Classification Metrics… Confusion Matrix

- There may be differential costs of making errors for different classes.
- For example, an incorrect medical diagnosis may be more costly than a false positive!
- So we need high confidence predictions only.
- Therefore, we can define other evaluation metrics based on a confusion matrix:

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

- Precision:
  - is the fraction of true positive (TP) samples that have been predicted positive over the total amount of predicted positive samples

Classification Metrics… Confusion Matrix

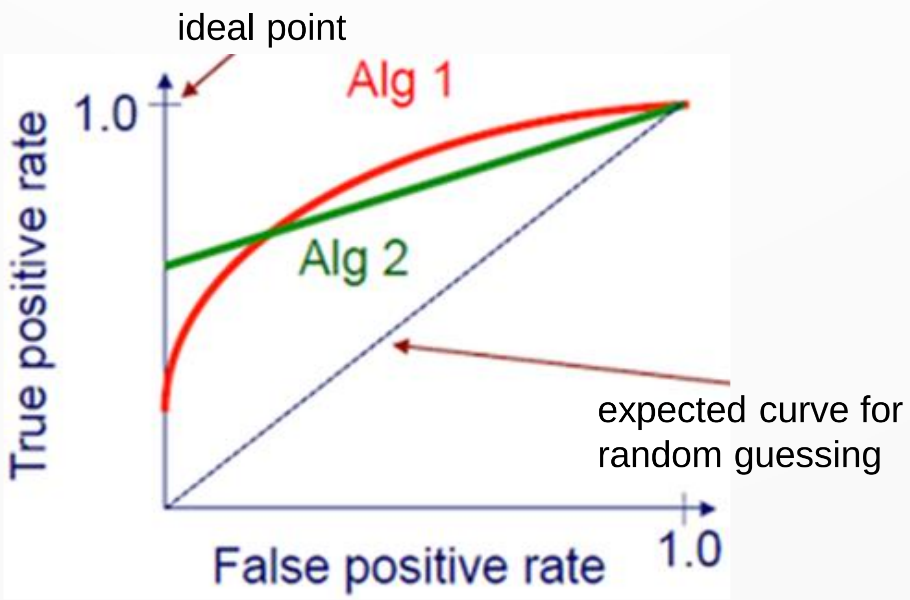
- Recall or True Positive Rate (TPR): $recall = \frac{TP}{TP \; + \; FN}$
  - is the fraction of true positive (TP) samples that have been predicted positive over the total amount of positive samples
- False positive rate (FPR): $FPR = \frac{FP}{TN \; + \; FP}$
  - is the fraction of false predicted positive (FP) samples over the total amount of negative samples

Classification Metrics…
ROC Curve

- Receiver Operating Characteristics (ROC) curve depicts the trade-off between the true positive rate and false positive rate.
    - ROC curve is especially useful for domains with imbalanced class distribution and unequal classification error costs.

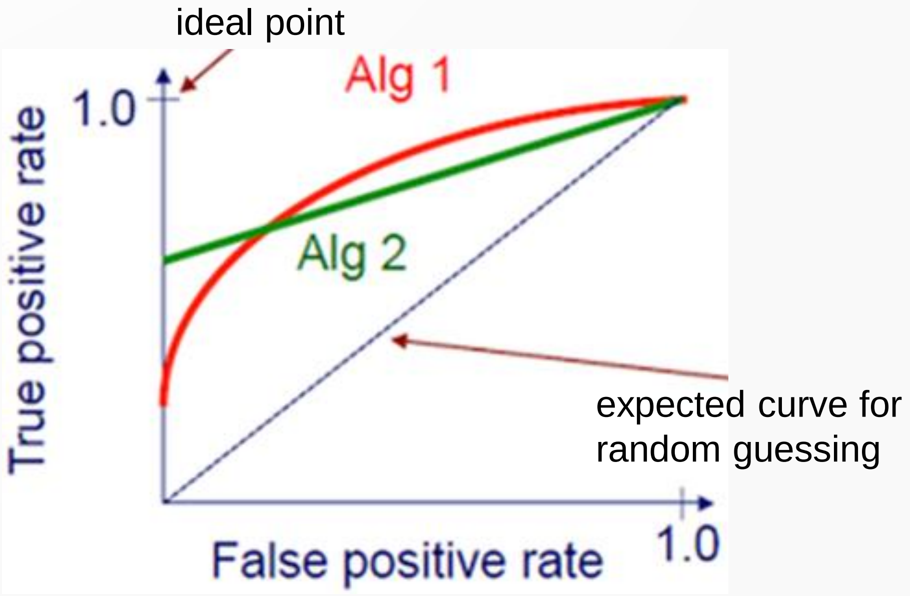
Classification Metrics…
ROC Curve

- The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.
  - TPR is also termed as Sensitivity
  - FPR is termed as 1-Specificity
- This has to be done to depict relative trade-offs between benefits (true positives) and costs (false positives).



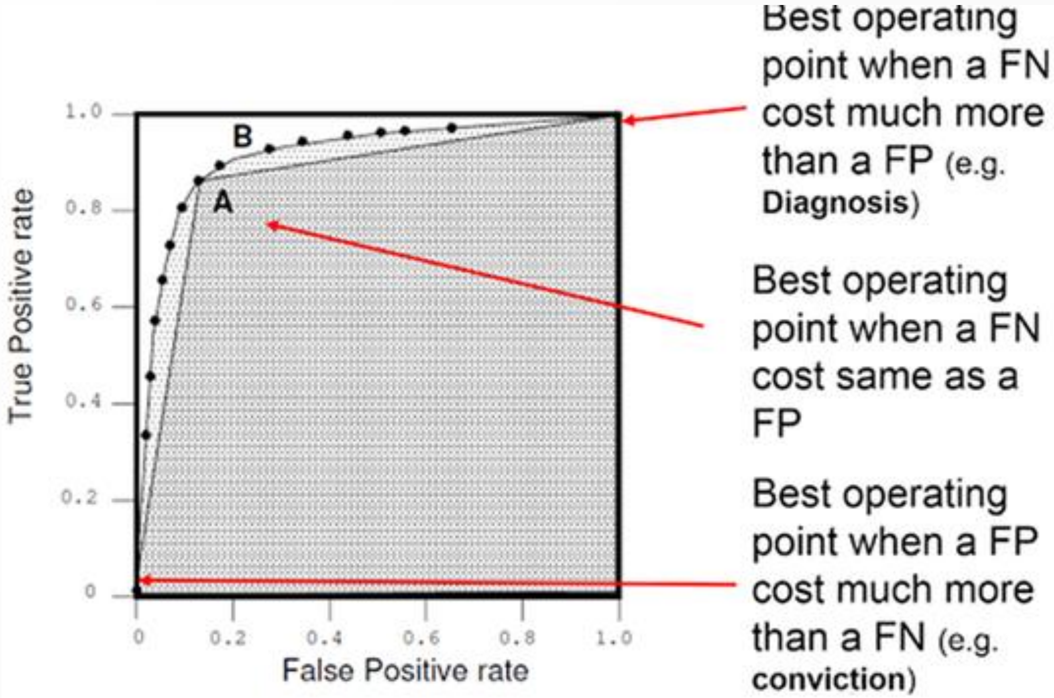
Classification Metrics…
ROC Curve

- As you can see in the figure below, different methods can work better in different parts of ROC space.
- There are two algorithms like Alg 1 and Alg 2 in the figure.
- The Alg 2 is good in the sense that it can give you high true positive rate while keeping the false positive rate low.
- in Alg 1, if it is allowed to incur more false positive rate, then Alg 1 can give us better higher true positive rate too.



Classification Metrics…
ROC Curve

- Lets say we are designing a classifier for a medical diagnosis.
- In this case we probably do not mind false positives!
  - since missing positive occurrence in detection
    of diseases are extremely costly.
- But, there can be situations where we do mind the false positive rate.
  - A good example of that could be in conviction for a crime. You do not want to waste someone's life with a false positive decision!



Classification Metrics… ROC Curve

- There are useful statistics that can be calculated via ROC curve, like the Area Under the Curve (AUC) and the Youden Index.
  - How well the model predicts and the optimal cut point for any given model (under specific circumstances).
  - AUC is used to summarize the ROC curve using a single number.
  - The higher the value of AUC, better performing is the classifier!
  - A random classifier has an AUC of 0.5.

Regression Metrics: Mean Square Error

- What are the ways of measuring regression performance?
- To measure how close the predictions are to the true target values, Mean Square Error (MSE) is a popular measure.
- MSE is defined as: $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Derived from MSE, Root Mean Square Error (RMSE) is also popular and is computed as: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
- Clearly, the lower the MSE of a model, the better its performance.
- Similar to MSE, Mean Absolute Error (MAE) is defined as: $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
- Due to using 1−norm of the error, MAE is robust to outliers in the test set.
  - Similar to MSE and RMSE, the lower the MAE of a model, the better its performance.

Regression Metrics: Explained Variance R^2

- This measure is known by many names including:
  - R-square, Explained variance, and coefficient of determination.
- R-square is measured as the percentage of target variation that is explained by the model.

$$R^2 = \frac{Variance\ Explained\ by\ the\ model}{Total\ variance}$$

- For linear regression with bias term, R-square is the square of the correlation between the target values and the predicted target values.
- Unlike the other introduced metrics,
  - the higher the R-square of a model, the better its performance.

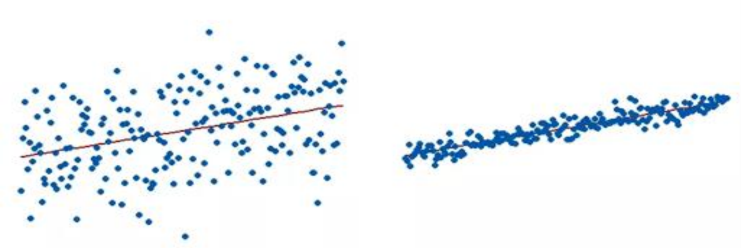
Regression Metrics: Explained Variance

R^2

- R-squared is always between 0 and 100%
  - 0% represents a model that does not explain any of the variation in the response variable around its mean.
  - The mean of the dependent variable predicts the dependent variable as well as the regression model.
  - 100% represents a model that explains all of the variation in the response variable around its mean.

Regression Metrics: Explained Variance R^2

- Consider the figure as illustration of regression in two cases.



- The R-squared for the regression model on
the left is ≤20%, and for the model on the right it is ≥80%.
- When a regression model accounts for more of the variance, the data points are closer to the regression line.

Thank You.