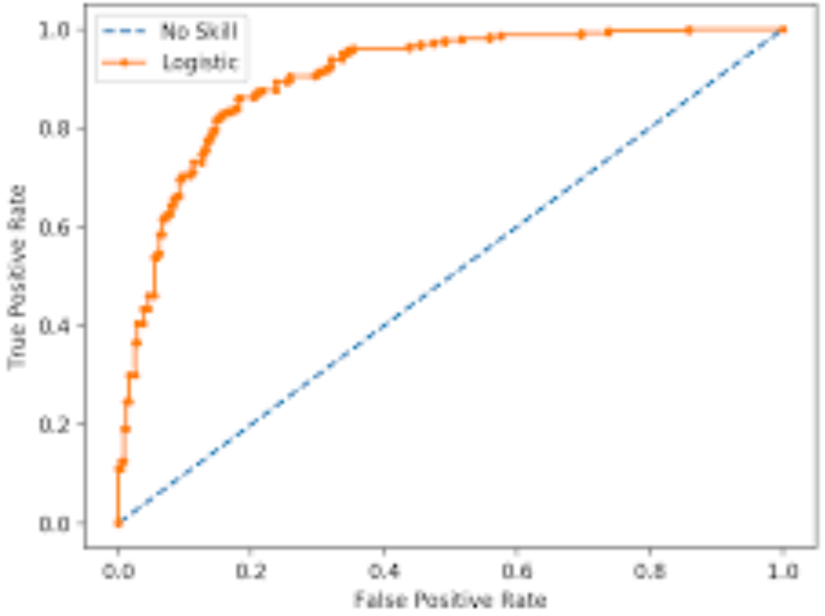
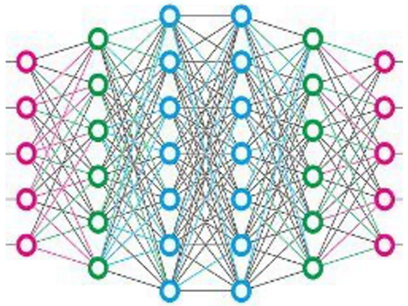
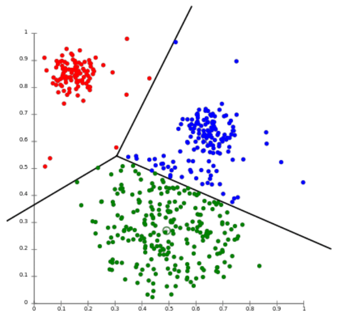
CS 445
Introduction to Machine Learning

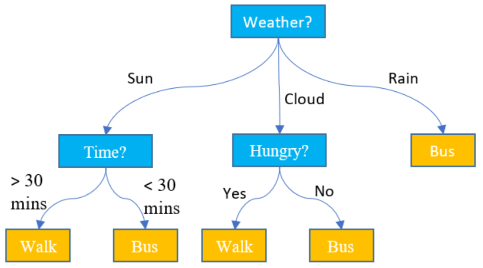
Imbalanced Classes
and
Comparing Classifiers with ROC Curves

Instructor: Dr. Kevin Molloy









Learning Objectives

- Define class imbalance
- Utilize sampling and synthetic sample generation for addressing class imbalances
- Costs of incorrect classification

Class Imbalance

Problems where the class occurrences are skewed:

- Credit card fraud
- Network intrusion detection
- Medical testing

Challenges:

- How to evaluate a model (accuracy is not well suited)

		Predicted Class	
		True (class=1)	False (Class=0)
Actual Class	True (class=1)	f_{11}(TP)	f_{10}(FN)
	False (class=0)	f_{01} (FP)	f_{00}(TN)

Precision = $\frac{TP}{(TP+FN)}$ Recall = $\frac{TP}{(TP+FP)}$

- Precision is the percentage correct considering only the actual positive class

New Measures

		Predicted Class	
		True (class=1)	False (Class=0)
Actual Class	True (class=1)	f_{11}(TP)	f_{10}(FN) (type II error)
	False (class=0)	f_{01} (FP) (Type I error)	f_{00}(TN)

Precision = $\frac{TP}{(TP+FP)}$ Recall = $\frac{TP}{(TP+FN)}$ Specificity = $\frac{TN}{(TN+FP)}$

- **Precision** is the percentage of correctly identified examples considering all examples that were labeled as positive.
- **Recall** is the percentage of true positives over all actual positive examples in the dataset. This is sometimes called **sensitivity** or the **true positive rate (TPR).**
- **Specificity** is the percentage of correctly identified examples that are negative out of all examples that are truly negative. Also known as the **true negative rate (TNR)**.

New Measures

		Predicted Class	
		True (class=1)	False (Class=0)
Actual Class	True (class=1)	f_{11}(TP)	f_{10}(FN) (type II error)
	False (class=0)	f_{01} (FP) (Type I error)	f_{00}(TN)

Precision = $\frac{TP}{(TP+FP)}$ Recall = $\frac{TP}{(TP+FN)}$ Specificity = $\frac{TN}{(TN+FP)}$

- **Precision** is the percentage of correctly identified examples considering all examples that were labeled as positive.
- **Recall** is the percentage of true positives over all actual positive examples in the dataset. This is sometimes called **sensitivity** or the **true positive rate (TPR).**
- **Specificity** is the percentage of correctly identified examples that are negative out of all examples that are truly negative. Also known as the **true negative rate (TNR)**.

F_1 measure = $\frac{2rp}{r+p}$ = $\frac{2 \cdot TP}{(2 \cdot TP+FP+ FN)}$ = $\frac{2}{\frac{1}{r}+\frac{1}{p}}$ Also known as the harmonic mean

Conveys a balance between precision and recall that is sensitive to the skew of the classes.

Creating a New Dataset

Dataset:

- 100 positive (+)
- 1,000 negative examples

Undersampling: Train by randomly sampling 100 of the negative values and using all 100 positive values.

Issues with this approach:

Creating a New Dataset

Dataset:

- 100 positive (+)
- 1,000 negative examples

Undersampling: Train by randomly sampling 100 of the negative values and using all 100 positive values.

Issues with this approach:

- **More important** examples of the over represented (negative) class could be omitted by random sampling
- Variance within the features may rise (since the number of examples is reduced)

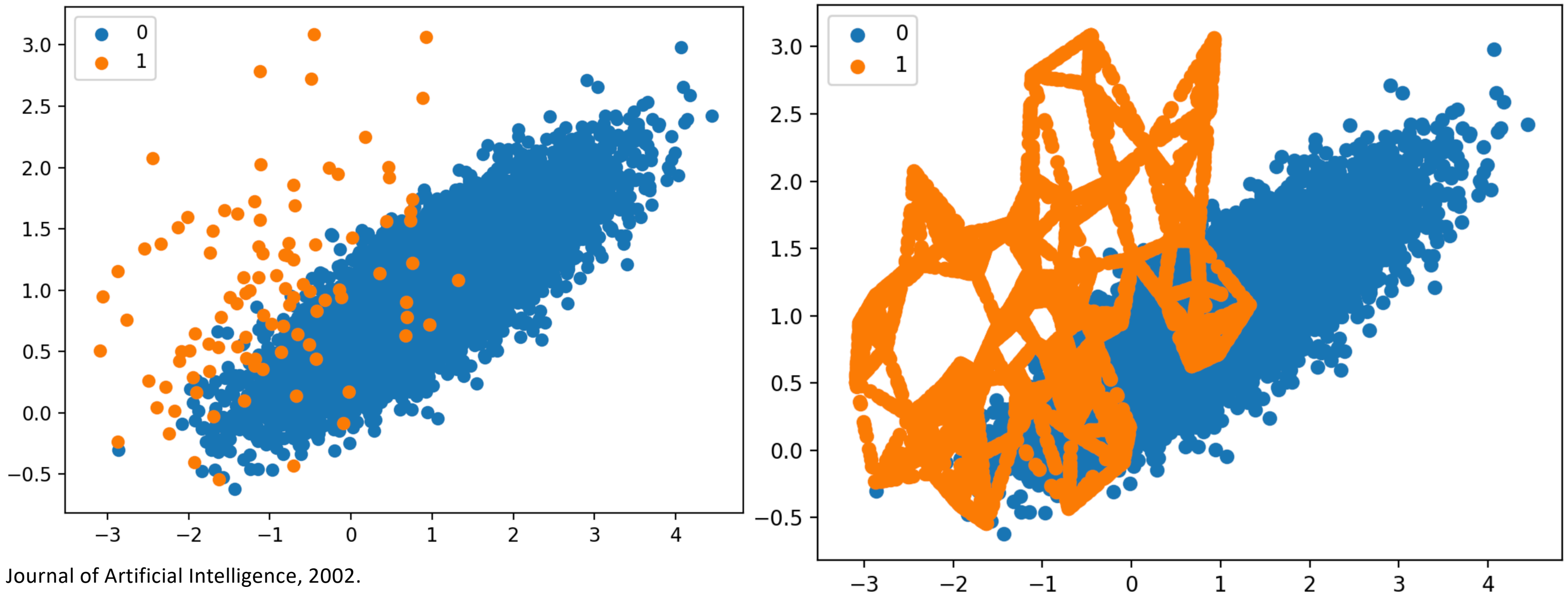
Question #1:

Pick a classifier (Decision Tree, KNN, Bayes, ANN): How does oversampling directly influence the specific classifier that you selected.

Generating More Data

Idea: Generate synthetic examples of the under represented class. Introducing the Synthetic Minority Oversampling Technique (SMOTE). The technique is as follows:

1. Select a positive example (x)
2. Determine x's k-nearest neighbors
3. Randomly select one of these neighbors (x_k)
4. Random generate a new example that lies on the line connecting x and x_k



Chawla et al. SMOTE: Synthetic Minority Oversampling Technique. Journal of Artificial Intelligence, 2002.

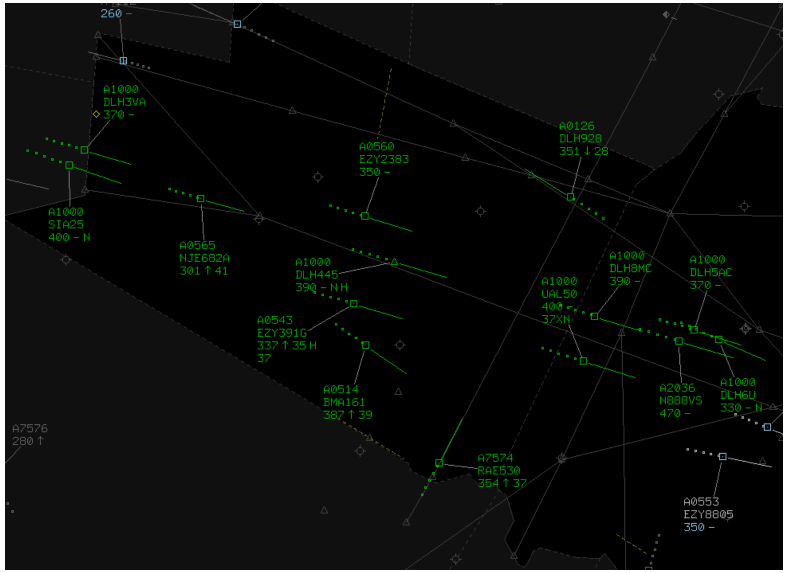
Plots from https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

Scoring

We can adapt the model to output a score, where higher scores indicate a strong tendency for the value to be in one class versus the other.

Idea: Find the optimal value where to set the scoring function. This can be found on the validation set (see 4.11.3 in the textbook).

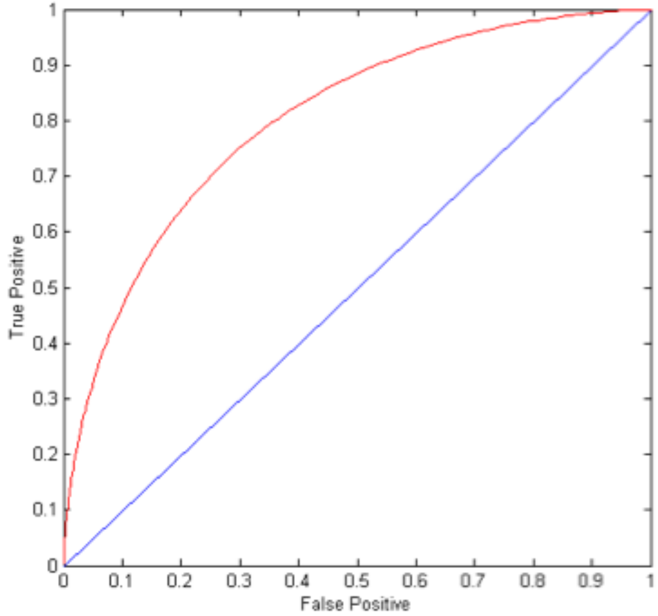
Evaluating All Scoring Thresholds



In some cases, there may be different costs associated with a FP than a FN.

Idea: Construct a method to evaluate all scores so that we can compare models across all levels of sensitivity.

The Receiver Operating Characteristic is a quantitative technique to evaluate the trade-off between detection rates (TPR) and the false alarm rate (FPR).

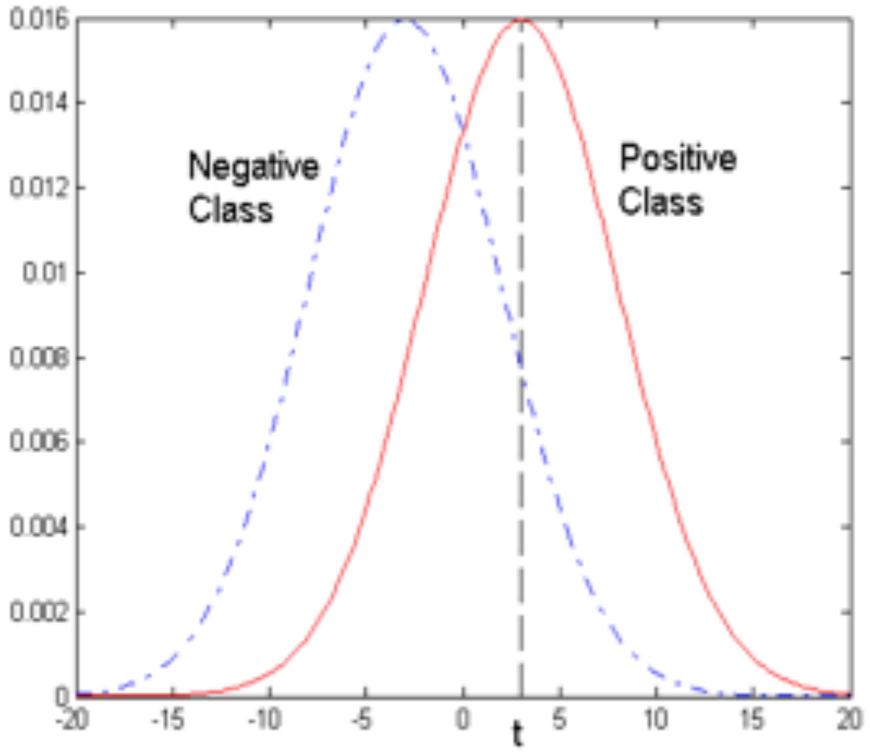


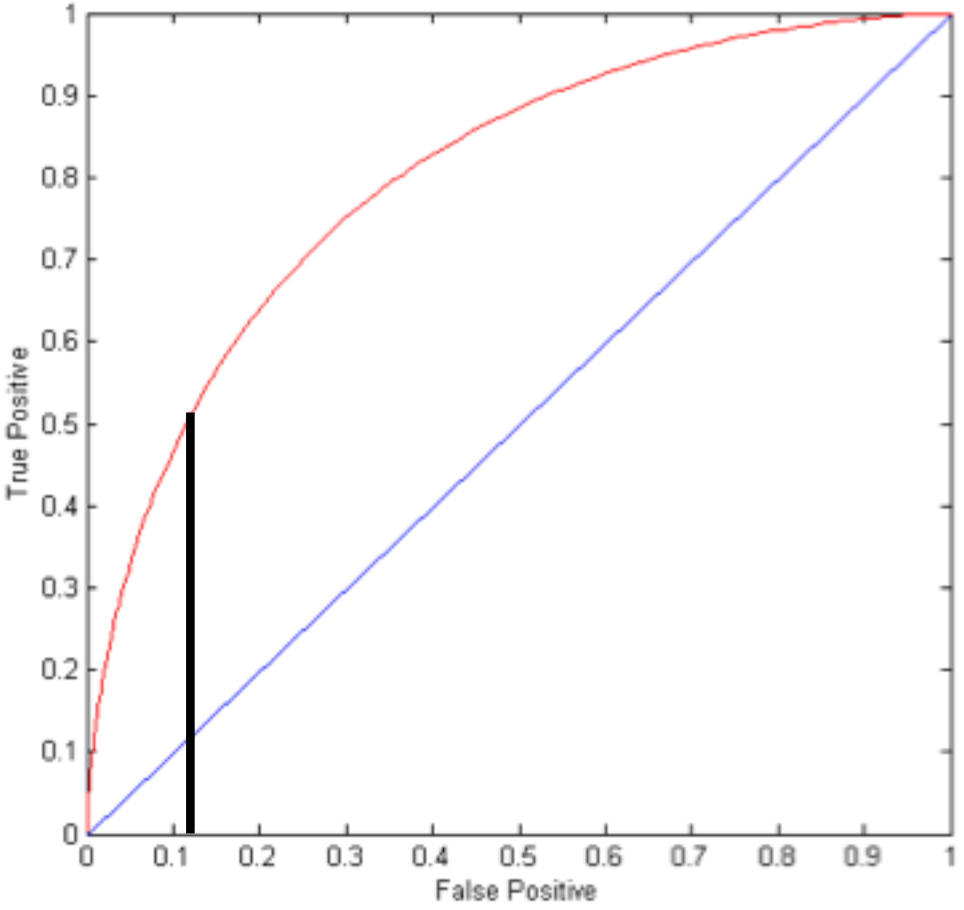
Precision/TPR = $\frac{TP}{(TP+FN)}$ FPR = $\frac{FP}{(FP+TN)}$

Notice that when we classify every instance as positive, both of these values are 1.

ROC (Receiver Operating Characteristic) Example

- 1-dimensional data set containing 2 classes (positive and negative)
- Points located at s(x) > t are classified as positive

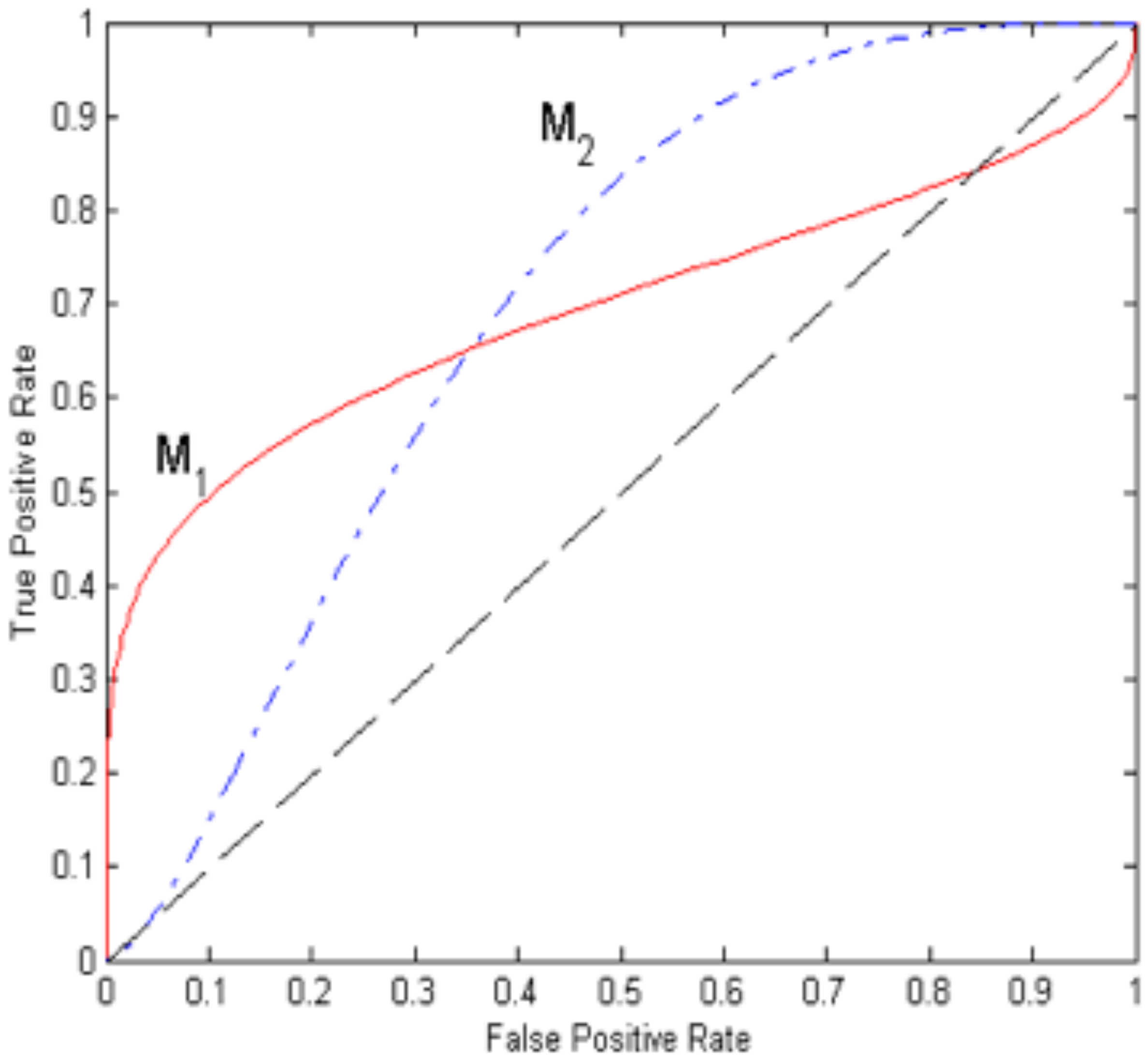




At threshold t:

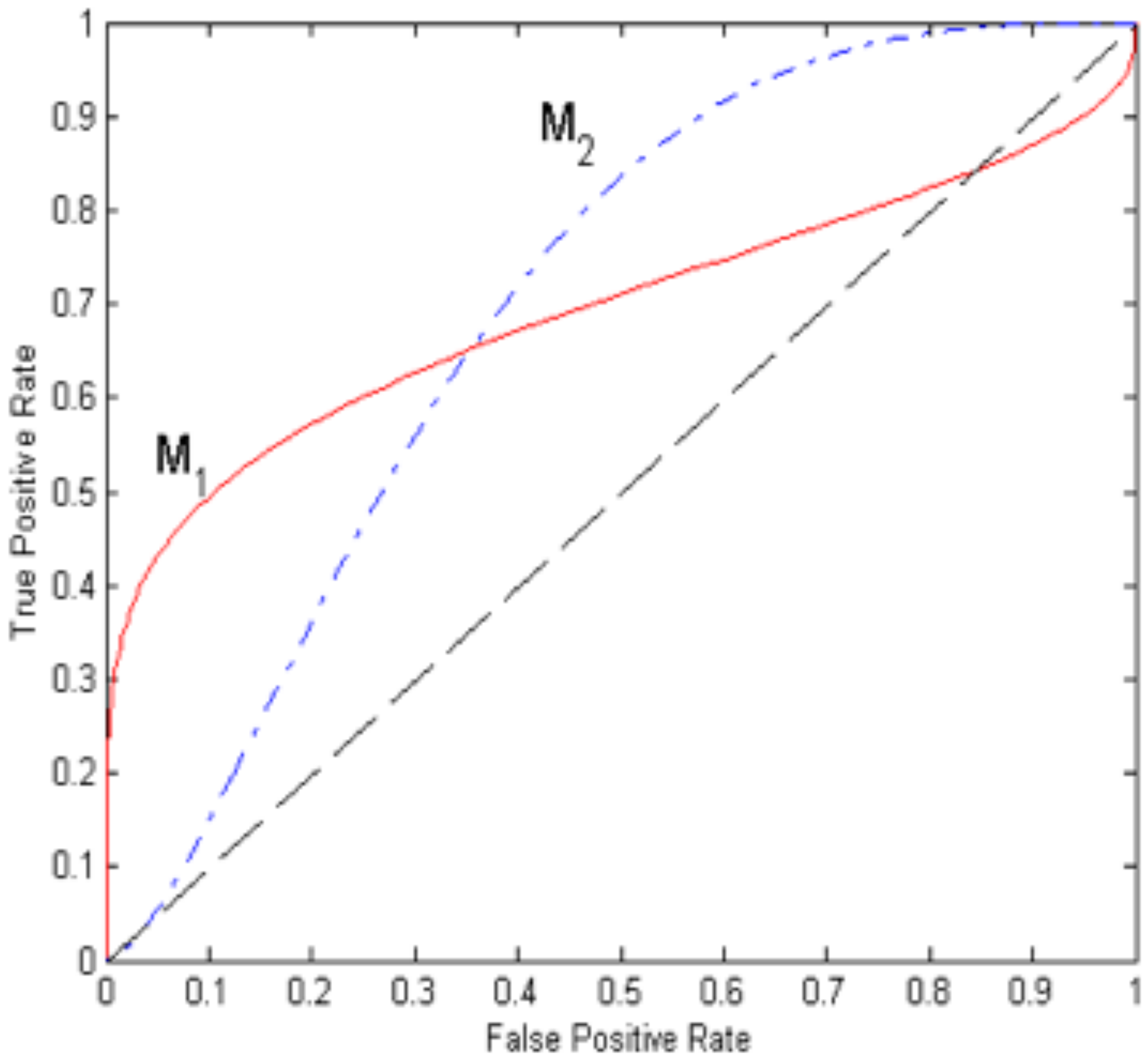
TP = 0.5, FN = 0.5, FP = 0.12 and TN = 0.88 Points located at s(x) > t are classified as positive.

Evaluating ROC Curves



Compare model performance:

Evaluating ROC Curves



Compare model performance:

M_1 is better if you need low false positive rates.

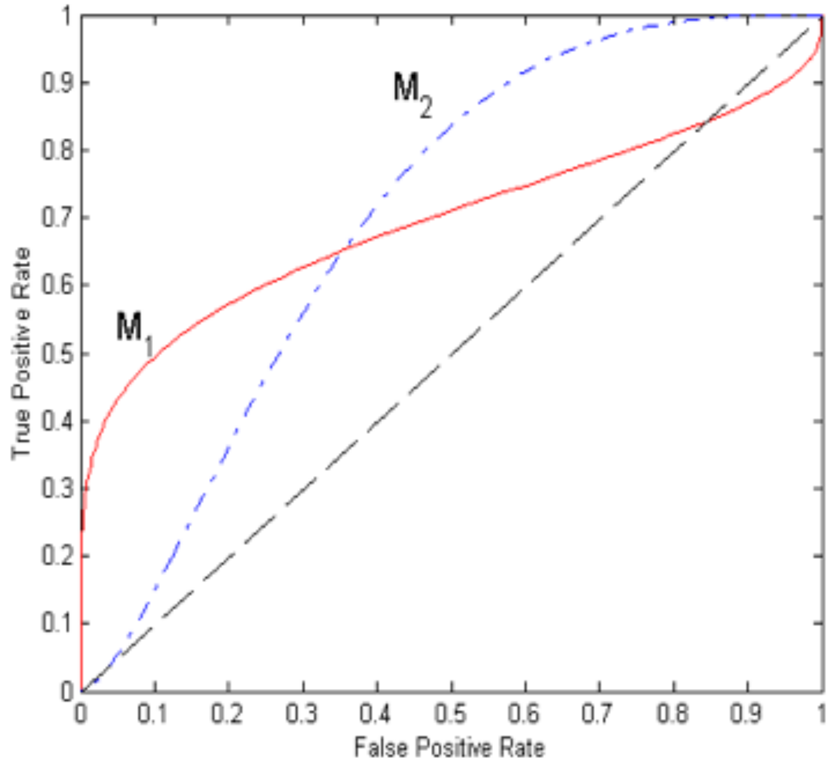
M_2 is better if higher false positive rates are OK.

Common quantitative metric is the area under the curve (AOC).

Constructing an ROC Curve

Requirement: Classifier must produce a score that resembles the posterior probability for each test instance (P (+ | A)).

1. Sort the instances according to P(+|A) in decreasing order
2. Apply threshold at each unique value of P(+ | A)
3. Count TP, FP, FN, TN
4. Plot FPR on X axis and TPR on y-axis



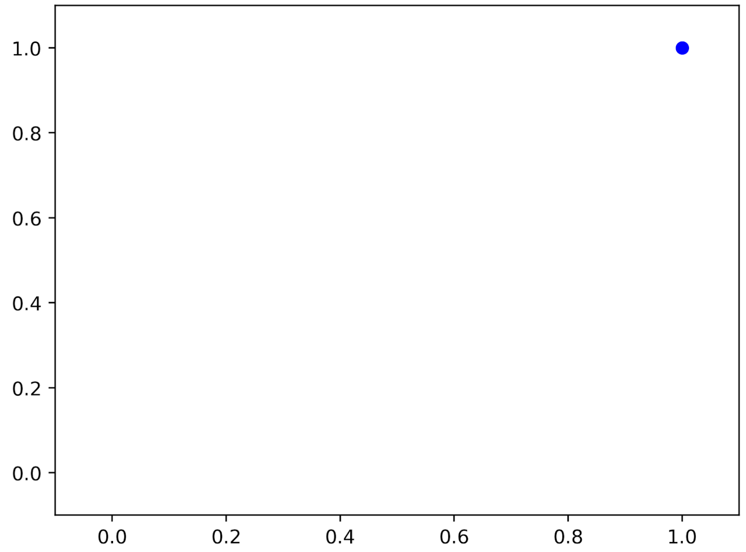
Constructing an ROC Curve

Class	+	-	+	-	-	-	+	-	+	+	
Thres >=	.25	.43	.53	.76	.85	.85	.85	.87	.93	.95	1.0
TP	5										
FP	5										
TN	0										
FN	0										
TPR	1										
FPR	1										

Precision/TPR = $\frac{TP}{(TP+FN)}$

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

Inst	P(+|A)	True Class
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+



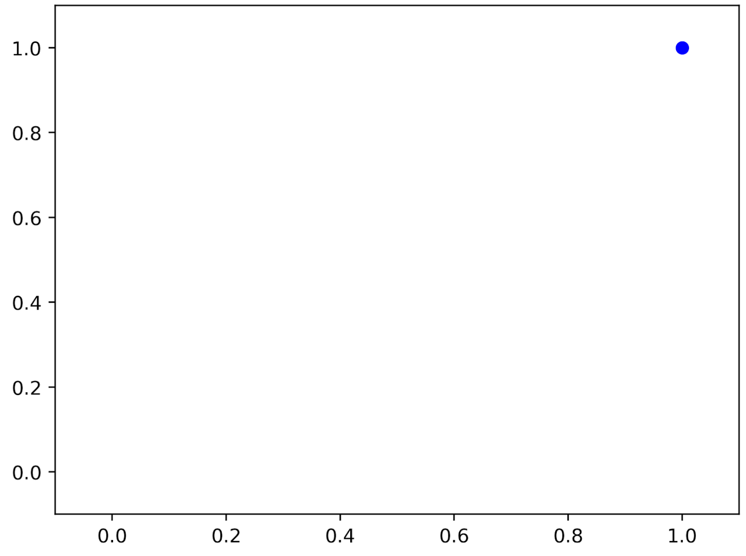
Constructing an ROC Curve

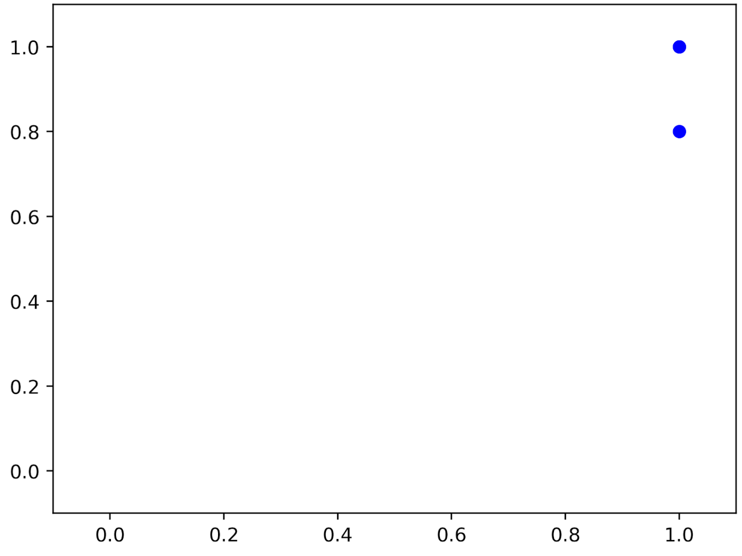
Class	+	-	+	-	-	-	+	-	+	+	
Thres >=	.25	.43	.53	.76	.85	.85	.85	.87	.93	.95	1.0
TP	5	4									
FP	5	5									
TN	0	0									
FN	0	0									
TPR	1	0.8									
FPR	1	1									

Precision/TPR = $\frac{TP}{(TP+FN)}$

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

Inst	P(+|A)	True Class
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+





Constructing an ROC Curve

Class	+	-	+	-	-	-	+	-	+	+	
Thres >=	.25	.43	.53	.76	.85	.85	.85	.87	.93	.95	1.0
TP	5	4									
FP	5	5									
TN	0	0									
FN	0	0									
TPR	1	0.8									
FPR	1	1									

Precision/TPR = $\frac{TP}{(TP+FN)}$

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

Inst	P(+|A)	True Class
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+