Evaluating Performance I

Lecture 06

Supervised learning in practice

Preprocessing
Explore & prepare data

Data Visualization and Exploration

Identify patterns that can be leveraged for learning

Data Cleaning

- Missing data
- Noisy data
- Erroneous data

Scaling (Standardization)

Prepare data for use in scale-dependent algorithms.

Feature Extraction

Dimensionality reduction eliminates redundant information

Model training

Select models (hypotheses)

Select model options that may fit the data well. We'll call them "hypotheses".

Fit the model to training data

Pick the "best" hypothesis function of the options by choosing model parameters

Iteratively fine tune the model

Performance evaluation

Make a prediction on validation data

Metrics

Classification
Precision, Recall, F_1, ROC Curves (Binary), Confusion Matrices (Multiclass)

Regression
MSE, explained variance, R^2

Supervised learning in practice

Preprocessing
Explore & prepare data

Data Visualization and Exploration

Identify patterns that can be leveraged for learning

Data Cleaning

- Missing data
- Noisy data
- Erroneous data

Scaling (Standardization)

Prepare data for use in scale-dependent algorithms.

Feature Extraction

Dimensionality reduction eliminates redundant information

Model training

Select models (hypotheses)

Select model options that may fit the data well. We'll call them "hypotheses".

Fit the model to training data

Pick the "best" hypothesis function of the options by choosing model parameters

Iteratively fine tune the model

Performance evaluation

Make a prediction on validation data

Metrics

Classification
Precision, Recall, F_1, ROC Curves (Binary), Confusion Matrices (Multiclass)

Regression
MSE, explained variance, R^2

Choose 2

Good

Cheap

Fast

Modeling Considerations

Accuracy

Computational Efficiency

Interpretability

Supervised Learning Performance Evaluation

Regression

Classification

Binary

Receiver Operating Characteristic (ROC) curves

Multiclass

Confusion matrices

Common Metrics

Regression

- Mean squared error (MSE)
- Mean absolute error (MAE)
- R^2, coefficient of determination
- Adjusted R^2

Binary

- Classification accuracy
- True positive rate
- False positive rate
- Precision
- F_1 Score
- Area under the ROC curve (AUC)

Multiclass

- Classification accuracy
- Micro-averaged F_1 Score
- Macro-averaged F_1 Score

Regression: Mean Squared Error

The mean squared error (MSE)

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Absolute measure of performance

One of the most widely used loss / cost functions

Regression: Mean **Absolute** Error

The mean absolute error (MAE)

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Absolute measure of performance

Regression: R^2 Coefficient of determination

Proportion of the response variable variation explained by the model

Residual sum of squares
(variation in the residuals)

$$SS_{res} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Total sum of squares
(variation in the data)

$$SS_{tot} = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

R-squared

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Relative measure of performance

Regression: Adjusted R^2

Problem: R^2 increases with more predictor variables

Adjusted R squared:

$$R^2_{adj} = 1 - (1 - R^2)\frac{N-1}{N-p-1}$$

Adjusts R squared to account for the number of predictor variables

This value is always less than or equal to the unadjusted R squared

Types of classification error

False Positive

(Type I error)

False Negative

(Type II error)

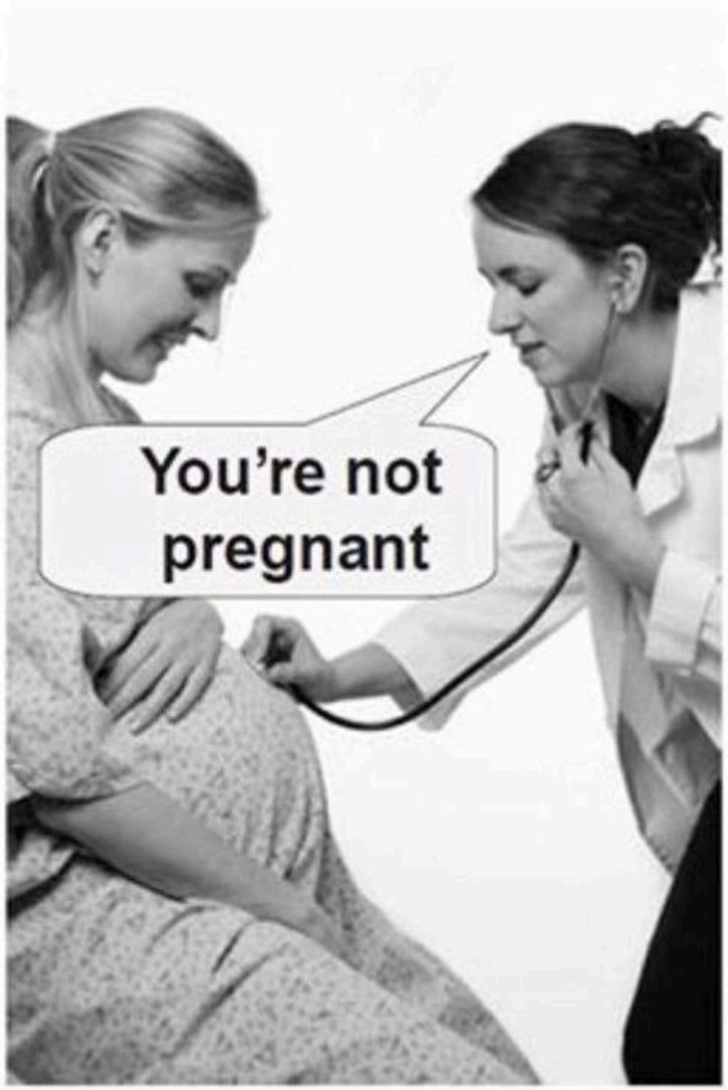


Image from: Ellis. *The Essential Guide to Effect Sizes*

Binary Classification

		Predicted Class, $\hat{y}$	
		Class 1 (target)	Class 0 (non-target)
True Class, y	Class 1 (target)	true positive	false negative
	Class 0 (non-target)	false positive	true negative

Type II Error
(missed target)

Type I Error

Binary Classification

Predicted Class, $\hat{y}$

True Class, y	Class 1 (target)	Class 0 (non-target)
Class 1 (target)	true positive	false negative
Class 0 (non-target)	false positive	true negative

True positive rate
Probability of detection, p_D
Sensitivity
Recall

$$\frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$

How many targets (Class 1) were correctly classified as targets?

Binary Classification

		Predicted Class, $\hat{y}$	
		Class 1 (target)	Class 0 (non-target)
True Class, y	Class 1 (target)	true positive	false negative
	Class 0 (non-target)	false positive	true negative

False positive rate
Probability of false alarm, p_{FA}

$$\frac{\text{false positive}}{\text{false positive} + \text{true negative}}$$

How many non-targets (Class 0) were incorrectly classified as targets?

Binary Classification

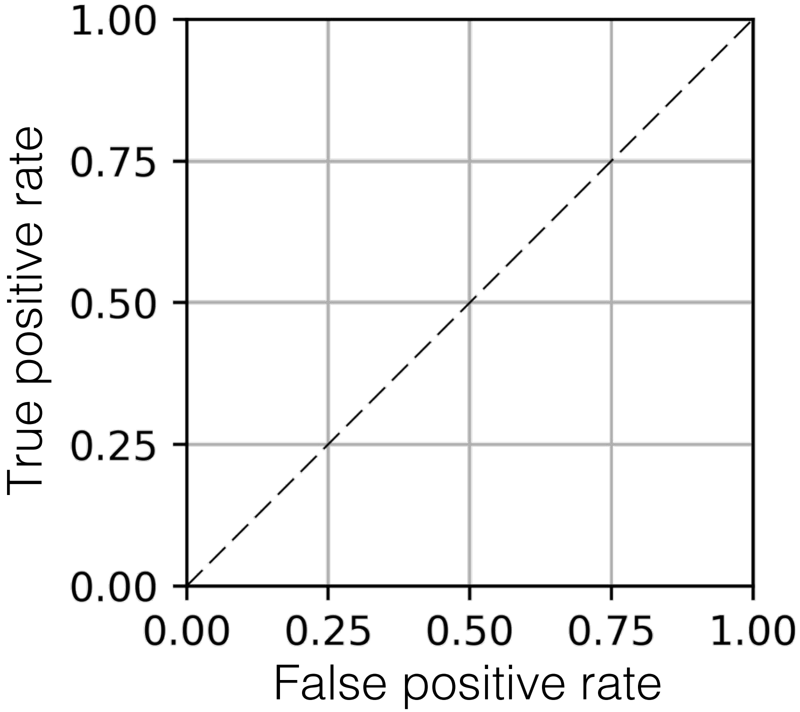
		Predicted Class, $\hat{y}$	
		Class 1 (target)	Class 0 (non-target)
True Class, y	Class 1 (target)	true positive	false negative
	Class 0 (non-target)	false positive	true negative

Precision

$$\frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$

How many of the predicted targets are targets?

ROC Curves



Estimate ($\hat{y}$)	True Class Label (y)	Classifier Confidence
?	1	1.40
?	1	0.95
?	0	0.80
?	1	0.60
?	0	-0.10

Classifier decision rule:

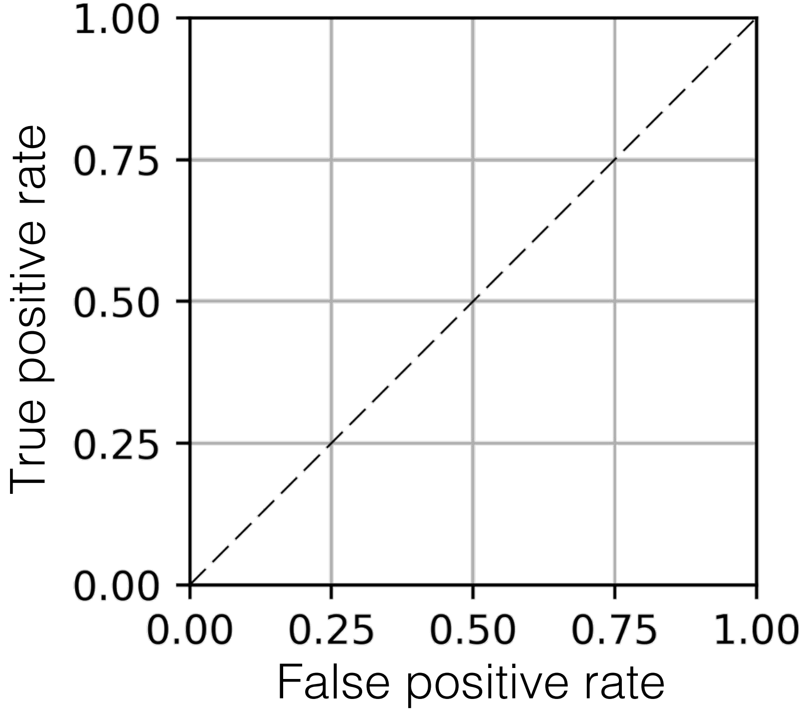
$$\hat{y} = \begin{cases} 1, & \text{confidence score} > \text{thresh} \\ 0, & \text{confidence score} \leq \text{thresh} \end{cases}$$

$$\frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$

$$\frac{\text{false positive}}{\text{false positive} + \text{true negative}}$$

Threshold	# True Positives	True Positive Rate	# False Positives	False Positive Rate

ROC Curves



True Class Label (y)	Classifier Confidence
1	1.40
1	0.95
0	0.80
1	0.60
0	-0.10

Classifier decision rule:

$$\hat{y} = \begin{cases} 1, & \text{confidence score} > \text{thresh} \\ 0, & \text{confidence score} \leq \text{thresh} \end{cases}$$

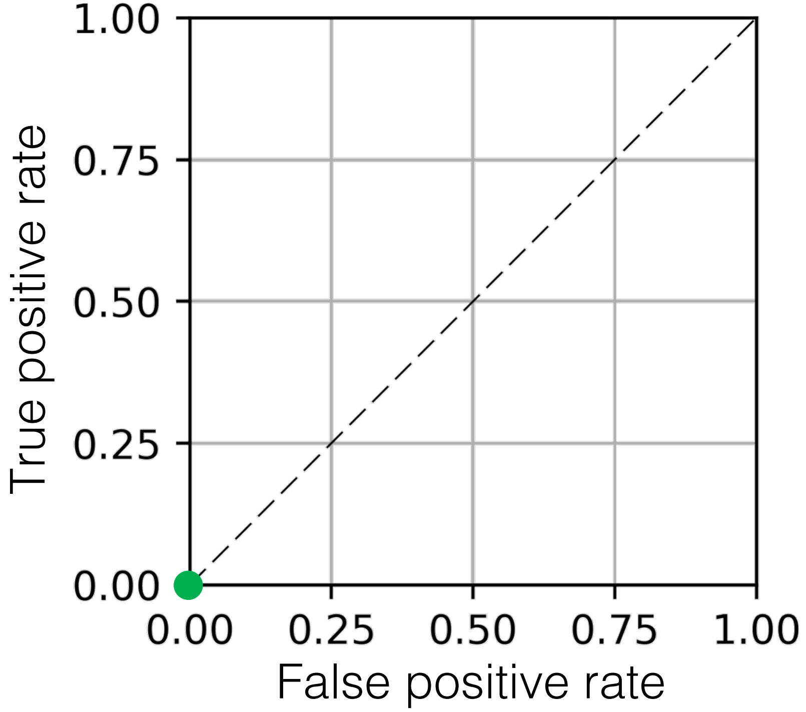
$$\frac{\text{true positive}}{\text{true positive} + \text{false negative}} \qquad \frac{\text{false positive}}{\text{false positive} + \text{true negative}}$$

Total Positives = 3

Total Negatives = 2

Threshold	# True Positives	True Positive Rate	# False Positives	False Positive Rate

ROC Curves



Classifier decision rule:

$$\hat{y} = \begin{cases} 1, & \text{confidence score} > \text{thresh} \\ 0, & \text{confidence score} \leq \text{thresh} \end{cases}$$

$$\frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$

$$\frac{\text{false positive}}{\text{false positive} + \text{true negative}}$$

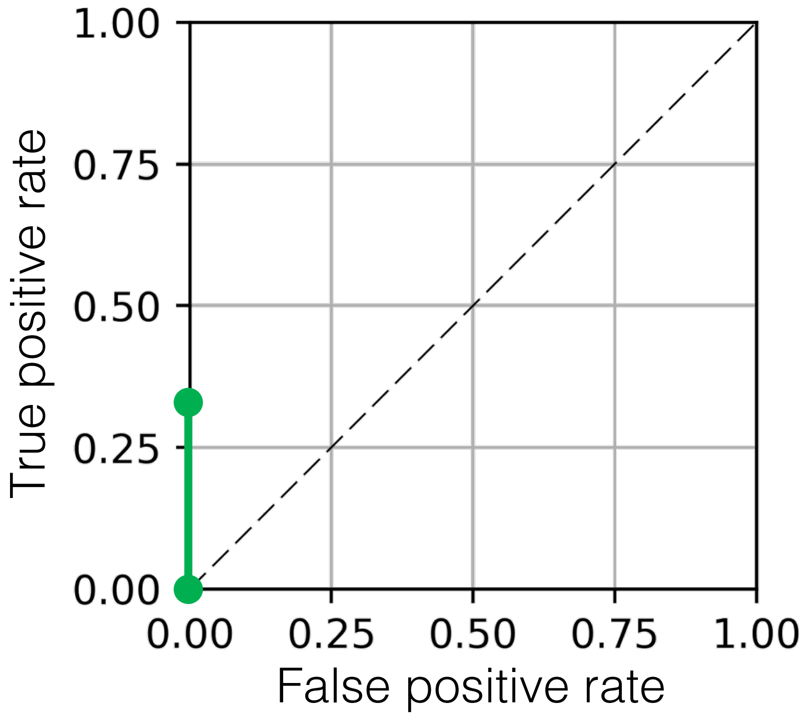
Total Positives = 3

Total Negatives = 2

Estimate ($\hat{y}$)	True Class Label (y)	Classifier Confidence
0	1	1.40
0	1	0.95
0	0	0.80
0	1	0.60
0	0	-0.10

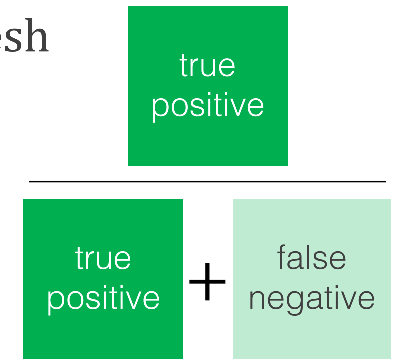
Threshold	# True Positives	True Positive Rate	# False Positives	False Positive Rate
∞	0	0	0	0

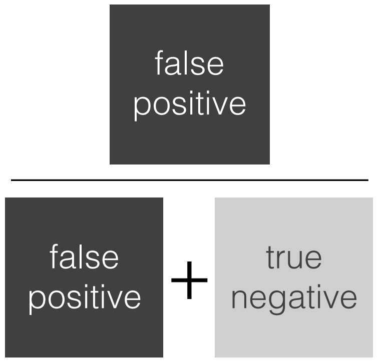
ROC Curves



Classifier decision rule:

$$\hat{y} = \begin{cases} 1, & \text{confidence score} > \text{thresh} \\ 0, & \text{confidence score} \leq \text{thresh} \end{cases}$$





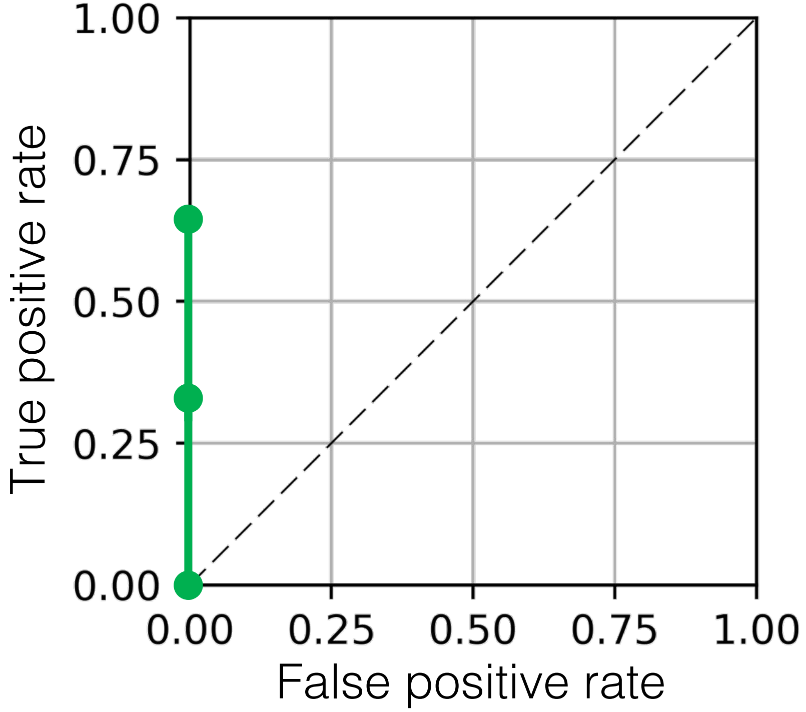
Total Positives = 3

Total Negatives = 2

Estimate ($\hat{y}$)	True Class Label (y)	Classifier Confidence
1	1	1.40
0	1	0.95
0	0	0.80
0	1	0.60
0	0	-0.10

Threshold	# True Positives	True Positive Rate	# False Positives	False Positive Rate
∞	0	0	0	0
1.0	1	0.333	0	0

ROC Curves



Classifier decision rule:

$$\hat{y} = \begin{cases} 1, & \text{confidence score} > \text{thresh} \\ 0, & \text{confidence score} \leq \text{thresh} \end{cases}$$

$$\frac{\text{true positive}}{\text{true positive} + \text{false negative}} \qquad \frac{\text{false positive}}{\text{false positive} + \text{true negative}}$$

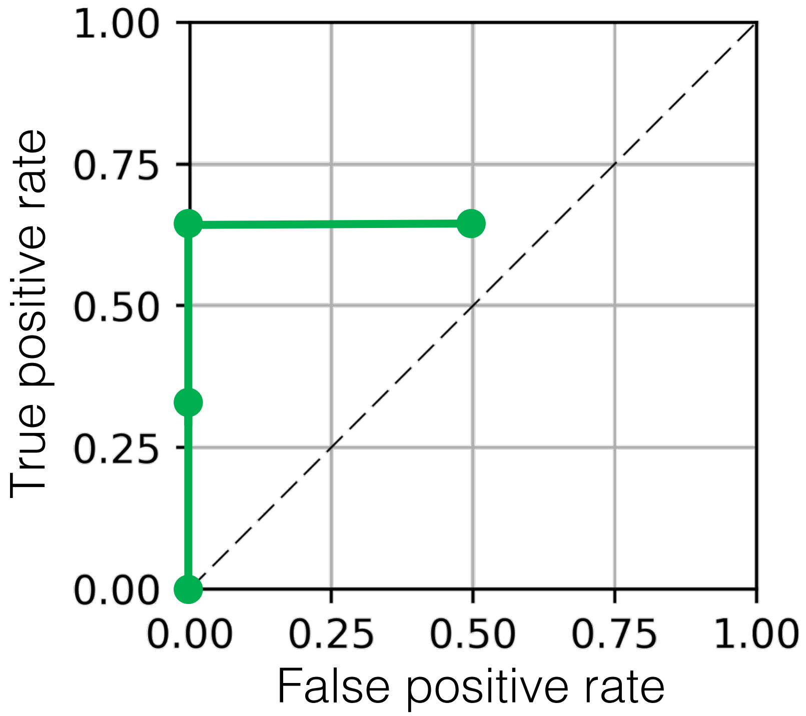
Total Positives = 3

Total Negatives = 2

Threshold	# True Positives	True Positive Rate	# False Positives	False Positive Rate
∞	0	0	0	0
1.0	1	0.333	0	0
0.9	2	0.667	0	0

Estimate (ŷ)	True Class Label (y)	Classifier Confidence
1	1	1.40
1	1	0.95
0	0	0.80
0	1	0.60
0	0	-0.10

ROC Curves



Classifier decision rule:

$$\hat{y} = \begin{cases} 1, & \text{confidence score} > \text{thresh} \\ 0, & \text{confidence score} \leq \text{thresh} \end{cases}$$

$$\frac{\text{true positive}}{\text{true positive} + \text{false negative}} \qquad \frac{\text{false positive}}{\text{false positive} + \text{true negative}}$$

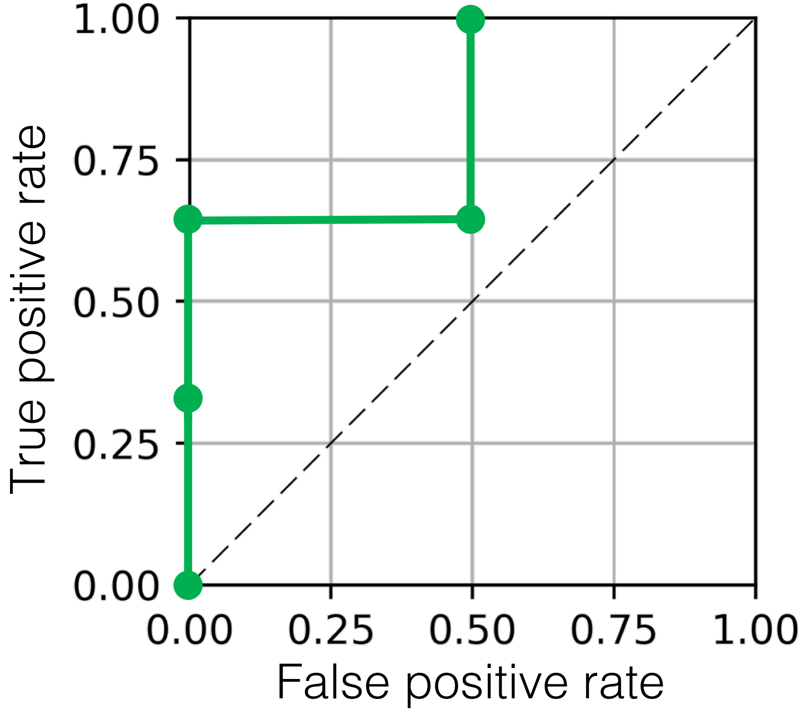
Total Positives = 3

Total Negatives = 2

Estimate ($\hat{y}$)	True Class Label (y)	Classifier Confidence
1	1	1.40
1	1	0.95
1	0	0.80
0	1	0.60
0	0	-0.10

Threshold	# True Positives	True Positive Rate	# False Positives	False Positive Rate
∞	0	0	0	0
1.0	1	0.333	0	0
0.9	2	0.667	0	0
0.7	2	0.667	1	0.5

ROC Curves



Classifier decision rule:

$$\hat{y} = \begin{cases} 1, & \text{confidence score} > \text{thresh} \\ 0, & \text{confidence score} \leq \text{thresh} \end{cases}$$

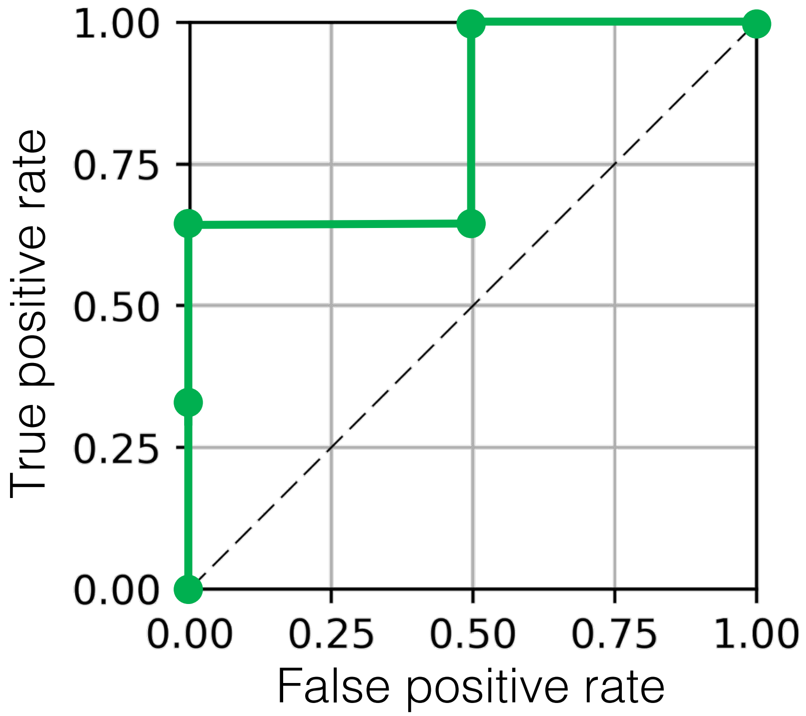
$$\frac{\text{true positive}}{\text{true positive} + \text{false negative}} \qquad \frac{\text{false positive}}{\text{false positive} + \text{true negative}}$$

Total Positives = 3 Total Negatives = 2

Estimate ($\hat{y}$)	True Class Label (y)	Classifier Confidence
1	1	1.40
1	1	0.95
1	0	0.80
1	1	0.60
0	0	-0.10

Threshold	# True Positives	True Positive Rate	# False Positives	False Positive Rate
∞	0	0	0	0
1.0	1	0.333	0	0
0.9	2	0.667	0	0
0.7	2	0.667	1	0.5
0.0	3	1	1	0.5

ROC Curves



Classifier decision rule:

$$\hat{y} = \begin{cases} 1, & \text{confidence score} > \text{thresh} \\ 0, & \text{confidence score} \leq \text{thresh} \end{cases}$$

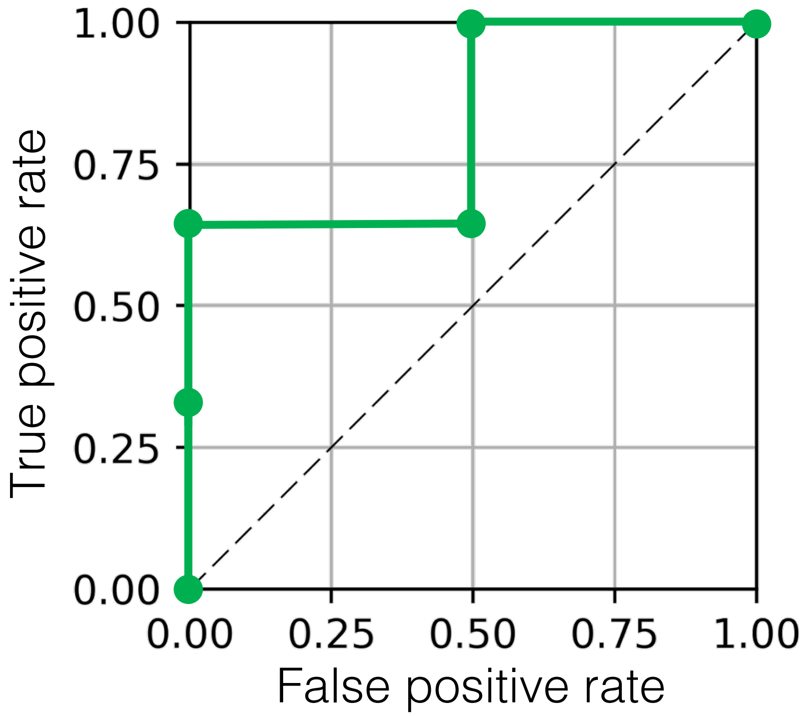
$$\frac{\text{true positive}}{\text{true positive} + \text{false negative}} \qquad \frac{\text{false positive}}{\text{false positive} + \text{true negative}}$$

Total Positives = 3 Total Negatives = 2

Estimate ($\hat{y}$)	True Class Label (y)	Classifier Confidence
1	1	1.40
1	1	0.95
1	0	0.80
1	1	0.60
1	0	-0.10

Threshold	# True Positives	True Positive Rate	# False Positives	False Positive Rate
∞	0	0	0	0
1.0	1	0.333	0	0
0.9	2	0.667	0	0
0.7	2	0.667	1	0.5
0.0	3	1	1	0.5
$-\infty$	3	1	2	1

ROC Curves



Estimate ($\hat{y}$)	True Class Label (y)	Classifier Confidence
1	1	1.40
1	1	0.95
1	0	0.80
1	1	0.60
1	0	-0.10

Classifier decision rule:

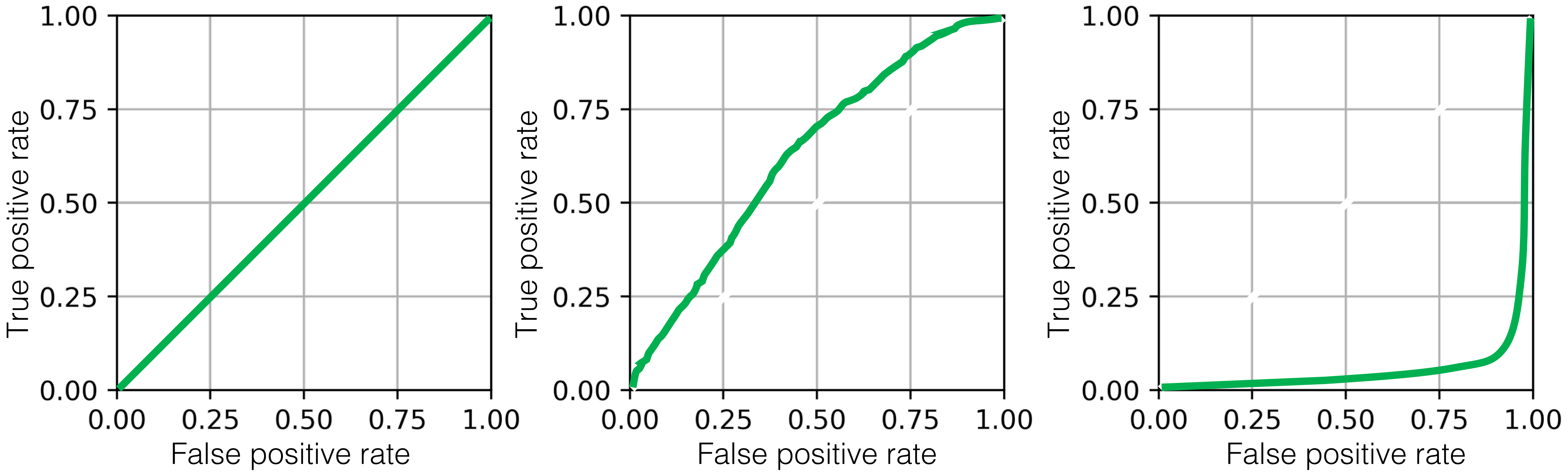
$$\hat{y} = \begin{cases} 1, & \text{confidence score} > \text{thresh} \\ 0, & \text{confidence score} \leq \text{thresh} \end{cases}$$

$$AUC = \left(\frac{2}{3}\right)\left(\frac{1}{2}\right) + (1)\left(\frac{1}{2}\right) = \frac{5}{6} \cong 0.833$$

Total Positives = 3 Total Negatives = 2

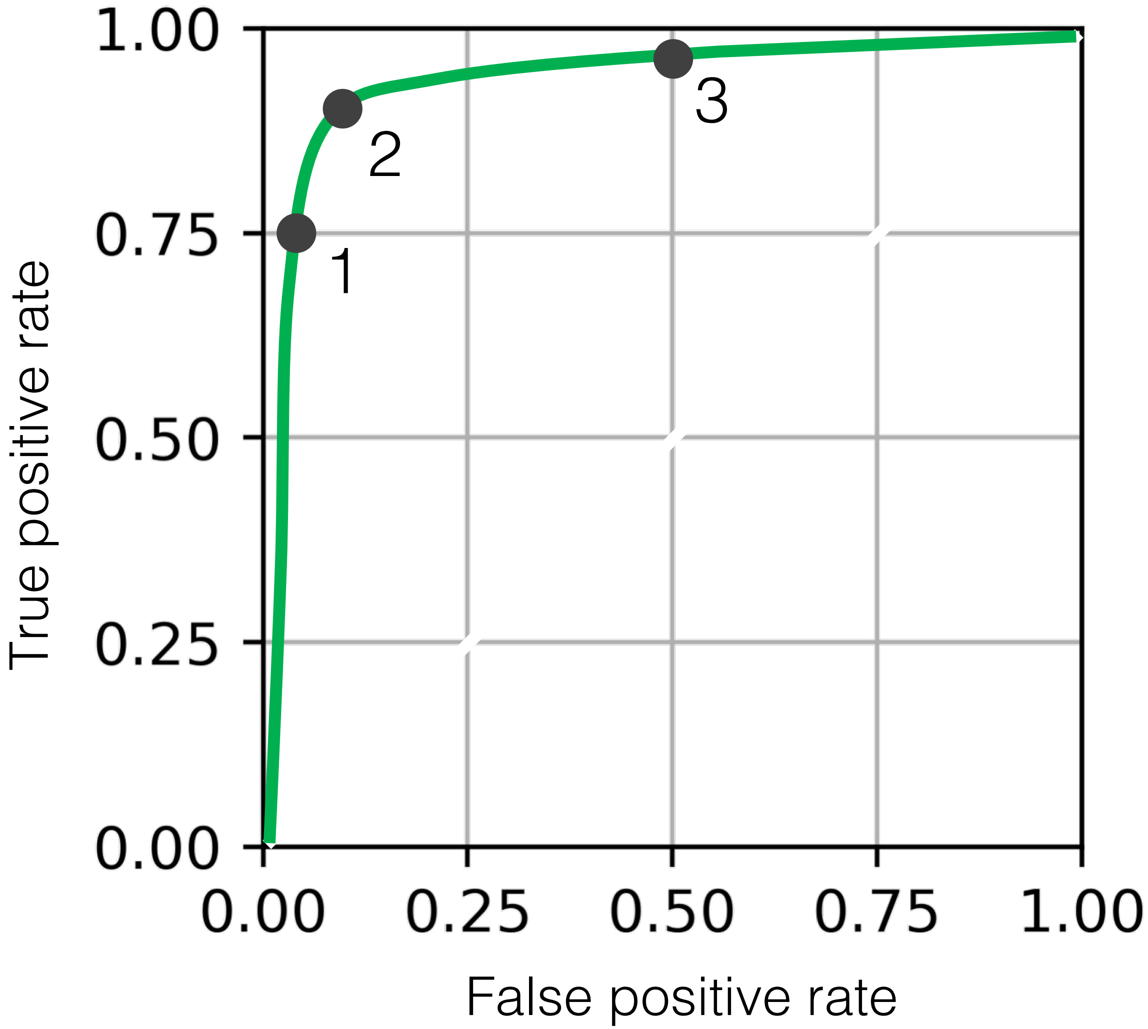
Threshold	# True Positives	True Positive Rate	# False Positives	False Positive Rate
∞	0	0	0	0
1.0	1	0.333	0	0
0.9	2	0.667	0	0
0.7	2	0.667	1	0.5
0.0	3	1	1	0.5
$-\infty$	3	1	2	1

ROC Curves: how do they compare?



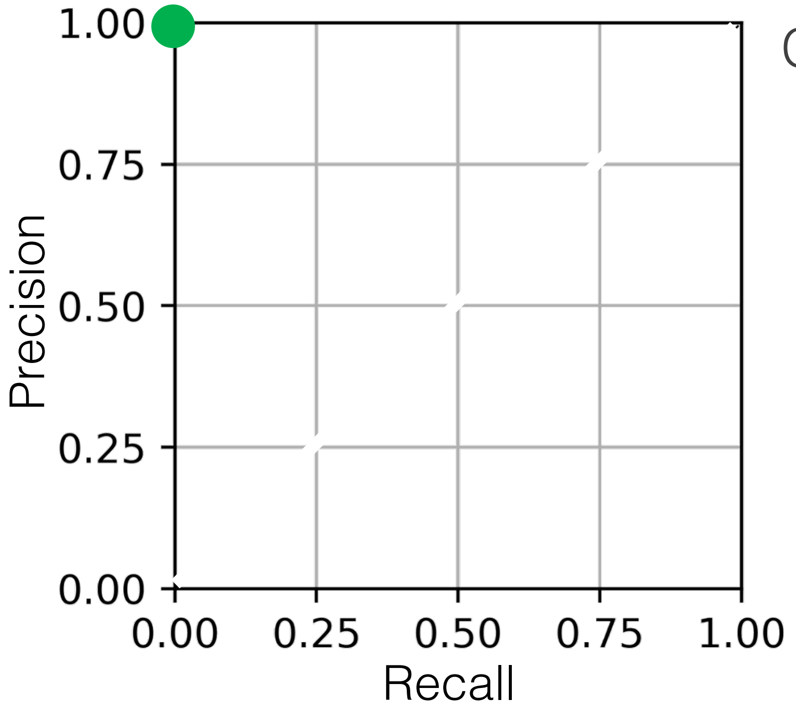
The model represented by this ROC curve is the most discriminative (but usually predicts incorrectly)

ROC Curves: where do we operate?



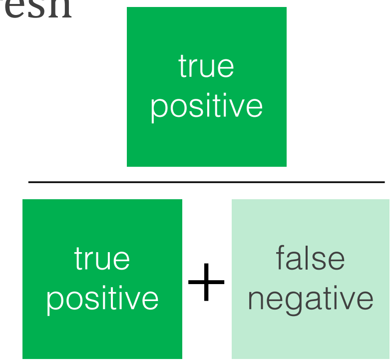
What does it mean to operate at a point on this curve?

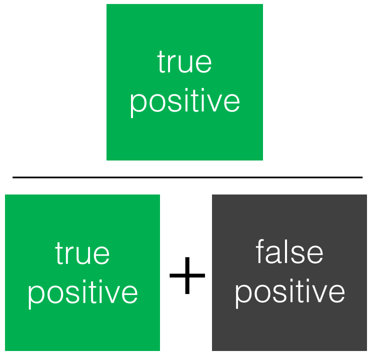
PR Curves



Classifier decision rule:

$$\hat{y} = \begin{cases} 1, \text{confidence score} > \text{thresh} \\ 0, \text{confidence score} \leq \text{thresh} \end{cases}$$



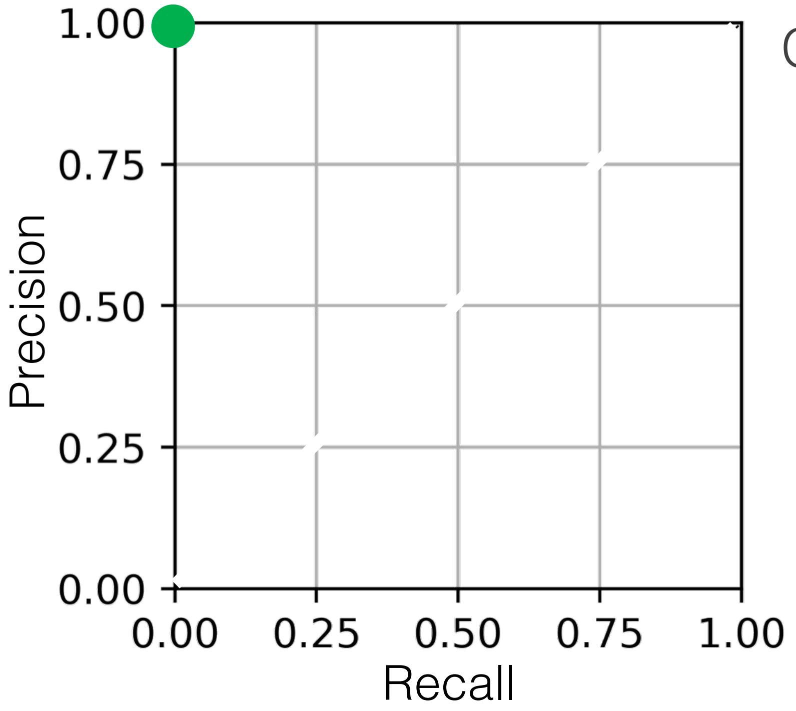


True Class Label (y)	Classifier Confidence
1	1.40
1	0.95
0	0.80
1	0.60
0	-0.10

Total Positives = 3 Total Negatives = 2

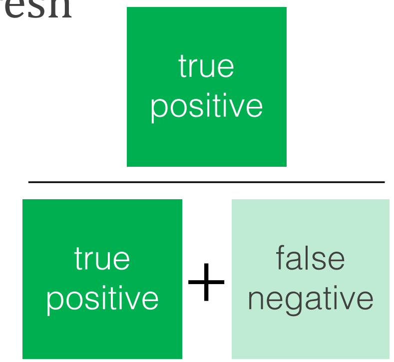
Threshold	# True Positives	Recall	# Predicted Positive	Precision

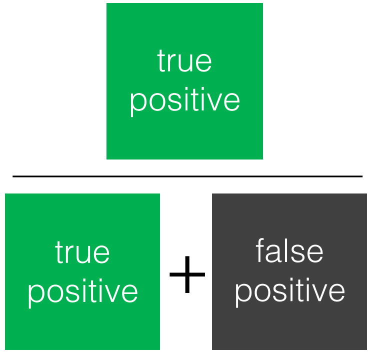
PR Curves



Classifier decision rule:

$$\hat{y} = \begin{cases} 1, \text{confidence score} > \text{thresh} \\ 0, \text{confidence score} \leq \text{thresh} \end{cases}$$



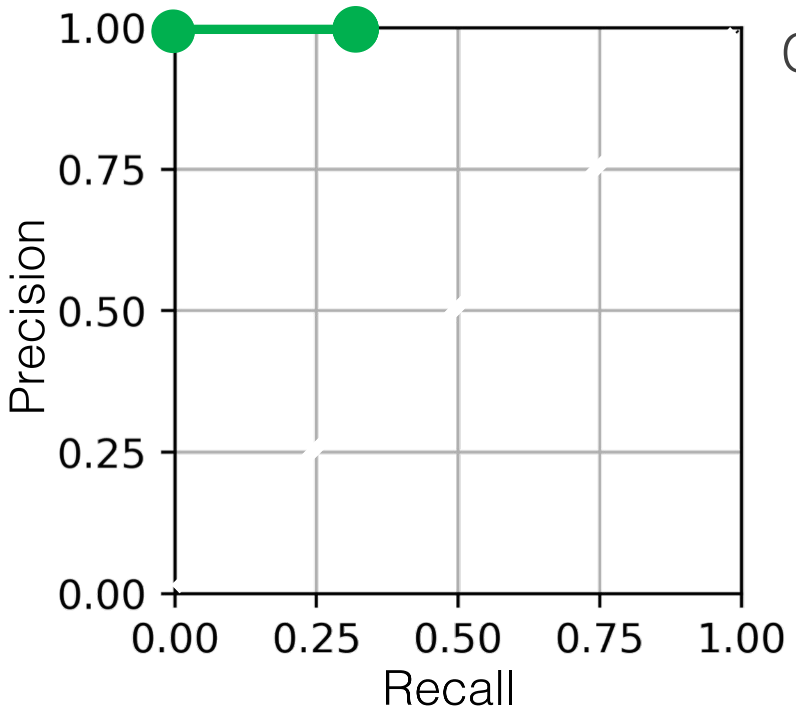


Total Positives = 3

Total Negatives = 2

Estimate ($\hat{y}$)	True Class Label (y)	Classifier Confidence
0	1	1.40
0	1	0.95
0	0	0.80
0	1	0.60
0	0	-0.10

Threshold	# True Positives	Recall	# Predicted Positive	Precision
∞	0	0	0	undefined



PR Curves

Classifier decision rule:

$$\hat{y} = \begin{cases} 1, \text{confidence score} > \text{thresh} \\ 0, \text{confidence score} \leq \text{thresh} \end{cases}$$

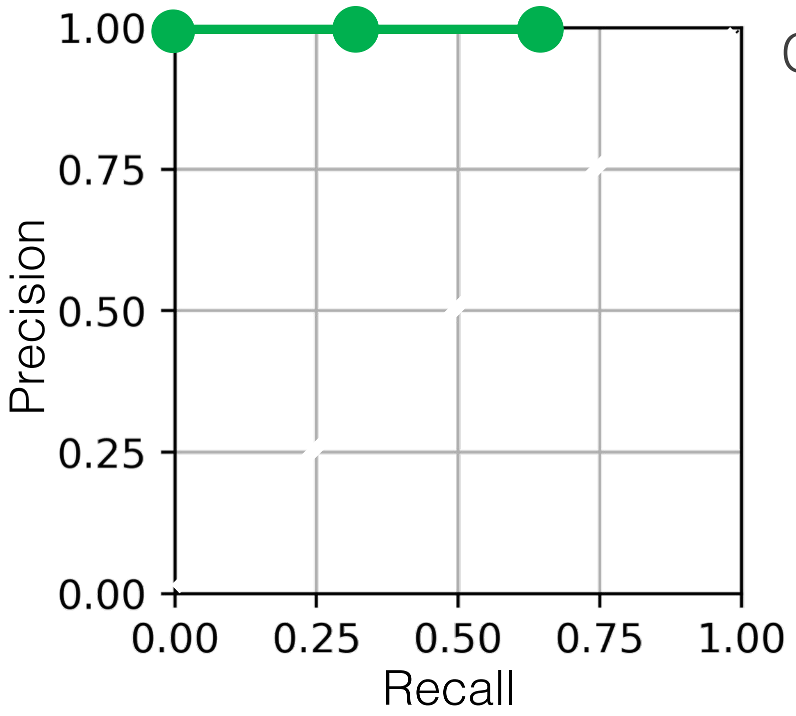
$$\frac{\text{true positive}}{\text{true positive} + \text{false negative}} \qquad \frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$

Estimate ($\hat{y}$)	True Class Label (y)	Classifier Confidence
1	1	1.40
0	1	0.95
0	0	0.80
0	1	0.60
0	0	-0.10

Total Positives = 3 Total Negatives = 2

Threshold	# True Positives	Recall	# Predicted Positive	Precision
∞	0	0	0	undefined
1.0	1	0.333	1	1

PR Curves



Classifier decision rule:

$$\hat{y} = \begin{cases} 1, \text{confidence score} > \text{thresh} \\ 0, \text{confidence score} \leq \text{thresh} \end{cases}$$

$$\frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$

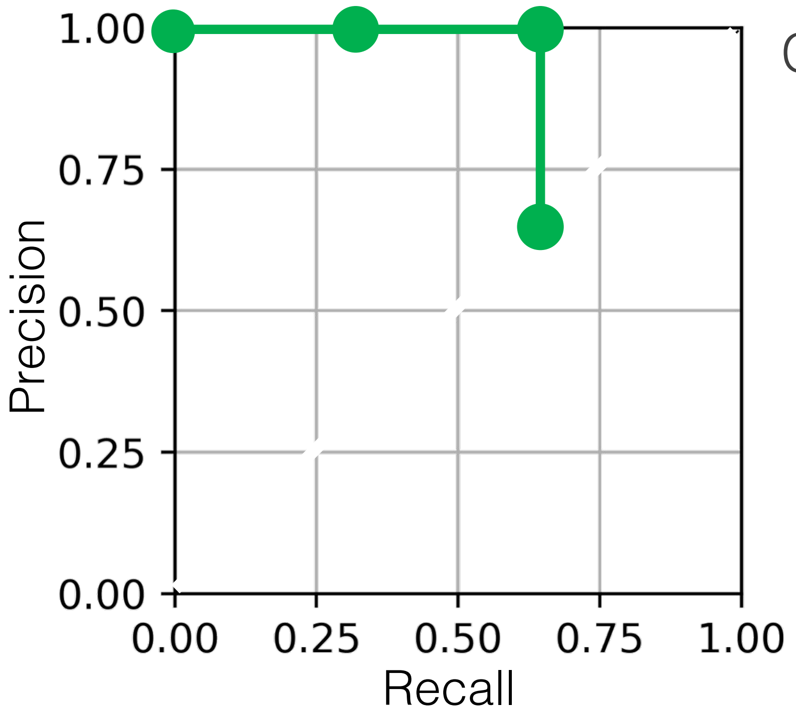
$$\frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$

Estimate ($\hat{y}$)	True Class Label (y)	Classifier Confidence
1	1	1.40
1	1	0.95
0	0	0.80
0	1	0.60
0	0	-0.10

Total Positives = 3 Total Negatives = 2

Threshold	# True Positives	Recall	# Predicted Positive	Precision
∞	0	0	0	undefined
1.0	1	0.333	1	1
0.9	2	0.667	2	1

PR Curves



Classifier decision rule:

$$\hat{y} = \begin{cases} 1, \text{confidence score} > \text{thresh} \\ 0, \text{confidence score} \leq \text{thresh} \end{cases}$$

$$\frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$

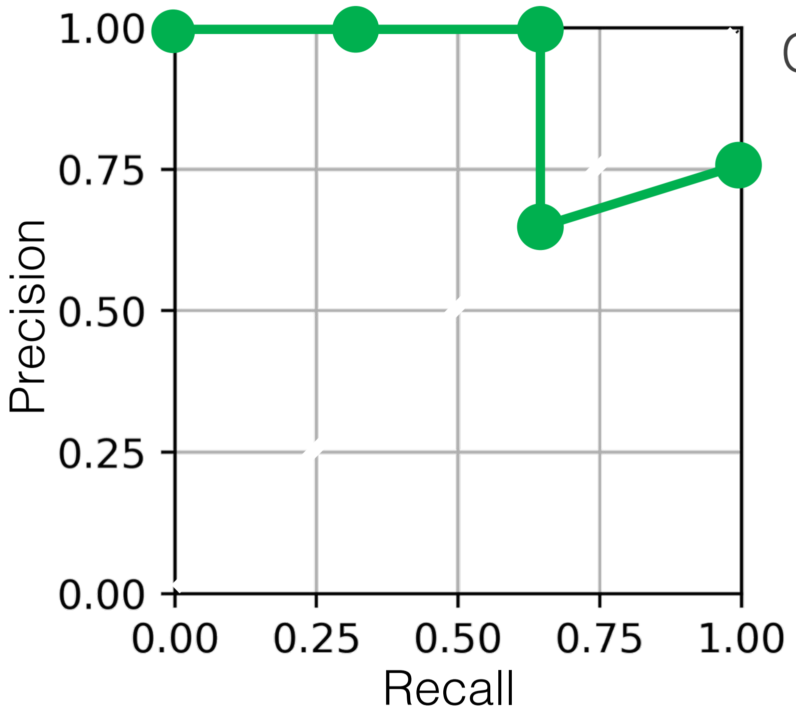
$$\frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$

Estimate ($\hat{y}$)	True Class Label (y)	Classifier Confidence
1	1	1.40
1	1	0.95
1	0	0.80
0	1	0.60
0	0	-0.10

Total Positives = 3 Total Negatives = 2

Threshold	# True Positives	Recall	# Predicted Positive	Precision
∞	0	0	0	undefined
1.0	1	0.333	1	1
0.9	2	0.667	2	1
0.7	2	0.667	3	0.667

PR Curves



Classifier decision rule:

$$\hat{y} = \begin{cases} 1, \text{confidence score} > \text{thresh} \\ 0, \text{confidence score} \leq \text{thresh} \end{cases}$$

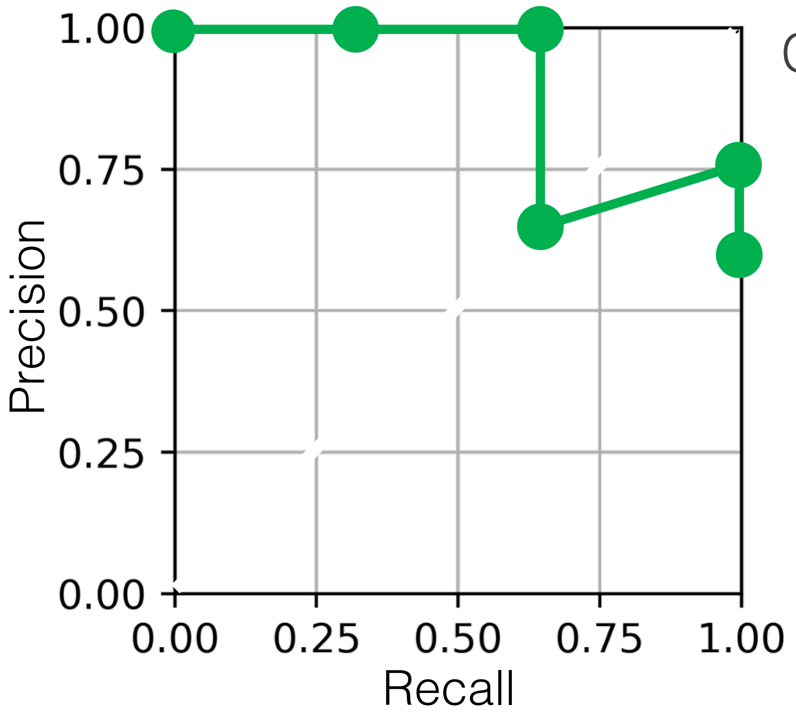
$$\frac{\text{true positive}}{\text{true positive} + \text{false negative}} \qquad \frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$

Estimate ($\hat{y}$)	True Class Label (y)	Classifier Confidence
1	1	1.40
1	1	0.95
1	0	0.80
1	1	0.60
0	0	-0.10

Total Positives = 3 Total Negatives = 2

Threshold	# True Positives	Recall	# Predicted Positive	Precision
∞	0	0	0	undefined
1.0	1	0.333	1	1
0.9	2	0.667	2	1
0.7	2	0.667	3	0.667
0.0	3	1	4	0.75

PR Curves



Classifier decision rule:

$$\hat{y} = \begin{cases} 1, \text{confidence score} > \text{thresh} \\ 0, \text{confidence score} \leq \text{thresh} \end{cases}$$

$$\frac{\text{true positive}}{\text{true positive} + \text{false negative}} \qquad \frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$

Estimate ($\hat{y}$)	True Class Label (y)	Classifier Confidence
1	1	1.40
1	1	0.95
1	0	0.80
1	1	0.60
1	0	-0.10

Total Positives = 3 Total Negatives = 2

Threshold	# True Positives	Recall	# Predicted Positive	Precision
∞	0	0	0	undefined
1.0	1	0.333	1	1
0.9	2	0.667	2	1
0.7	2	0.667	3	0.667
0.0	3	1	4	0.75
$-\infty$	3	1	5	0.6

Be wary of overall accuracy as sole metric

Case study 1

i	y_i	$\hat{y}_i$
1	**1**	**1**
2	**1**	**1**
3	**1**	**1**
4	**1**	**1**
5	**1**	**1**
6	**1**	**1**
7	**1**	**0**
8	**0**	**1**
9	**0**	**0**
10	**0**	**0**
11	**0**	**0**
12	**0**	**0**
13	**0**	**0**
14	**0**	**0**
15	**0**	**0**

Overall classification accuracy = 13/15 = **0.87**

ROC Curves measure the tradeoff between…

A False positive rate = 1/8 = **0.13**

B True positive rate (Recall) = 6/7 = **0.86**

PR Curves measure the tradeoff between…

B True positive rate (Recall) = 6/7 = **0.86**

C Precision= 6/7 = **0.86**

A $\dfrac{\text{false positive}}{\text{false positive} + \text{true negative}}$

B $\dfrac{\text{true positive}}{\text{true positive} + \text{false negative}}$

C $\dfrac{\text{true positive}}{\text{true positive} + \text{false positive}}$

Case study 2

i	y_i	$\hat{y}_i$
1	1	1
2	1	1
3	1	0
4	1	0
5	0	0
6	0	0
7	0	0
8	0	0
9	0	0
10	0	0
11	0	0
12	0	0
13	0	0
14	0	0
15	0	0

Overall classification accuracy = 13/15 = **0.87**

ROC Curves measure the tradeoff between…

A False positive rate = 0/11 = **0**

B True positive rate (Recall) = 2/4 = **0.5**

PR Curves measure the tradeoff between…

B True positive rate (Recall) = 2/4 = **0.5**

C Precision= 2/2 = **1**

A $\dfrac{\text{false positive}}{\text{false positive} + \text{true negative}}$

B $\dfrac{\text{true positive}}{\text{true positive} + \text{false negative}}$

C $\dfrac{\text{true positive}}{\text{true positive} + \text{false positive}}$

Case study 3

i	y_i	$\hat{y}_i$
1	**1**	**1**
2	**1**	**1**
3	**1**	**1**
4	**1**	**1**
5	**1**	**1**
6	**1**	**1**
7	**1**	**1**
8	**1**	**1**
9	**1**	**1**
10	**1**	**1**
11	**1**	**1**
12	**1**	**1**
13	**1**	**1**
14	**0**	**1**
15	**0**	**1**

Overall classification accuracy = 13/15 = **0.87**

ROC Curves measure the tradeoff between…

A False positive rate = 2/2 = **1**

B True positive rate (Recall) = 13/13 = **1**

PR Curves measure the tradeoff between…

B True positive rate (Recall) = 13/13 = **1**

C Precision= 13/15 = **0.87**

A $\dfrac{\text{false positive}}{\text{false positive} + \text{true negative}}$

B $\dfrac{\text{true positive}}{\text{true positive} + \text{false negative}}$

C $\dfrac{\text{true positive}}{\text{true positive} + \text{false positive}}$

Multiclass Classification: Confusion Matrix

Predicted Class, $\hat{y}$

True Class, y	Class 1	Class 2	Class 3	No. samples from class
Class 1	190	8	2	[200]
Class 2	1	5	4	[10]
Class 3	24	24	25	[73]

confusion matrix with number of samples

F_1-score

$$F_1 = 2\frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}}$$

Harmonic mean of precision and recall

$$= 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Generally:

$$F_\beta = (1 + \beta^2)\frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

β controls the relative weight of precision/recall

Multiclass F_1

Micro-average: Calculate precision and recall metrics globally by counting the total true positives, false negatives, and false positives
(average for the whole dataset)

Macro-average: Use the average precision and recall for each class label
(average of class-averages)