Evaluation Metrics for ML Performance

Guowei Wei
Department of Mathematics
Michigan State University

References:
Duc D. Nguyen’s lecture notes
Wikipedia

Introduction

- Use statistical values to evaluate the performance of a ML algorithm
- Compare the predictive power between different ML predictors
- We need to analyze two types of predictors, regressors and classifiers.

Evaluation Metrics for Regressors

- Root mean square error (RMSE)
- Pearson correlation (R_p)
- Spearman correlation (R_s)
- Kendall Tau (τ)

RMSE

- Assume we have M true labels

$$y_1, y_2, \dots, y_M$$

Our predictor gives M predicted labels

$$\hat{y}_1, \hat{y}_2, \dots, \hat{y}_M$$

(To avoid heavy notation we set

$y_i \equiv y^{(i)}$ and $\hat{y}_i \equiv \hat{y}^{(i)}$ in this lecture)

RMSE will measure the root mean square errors between predicted labels and the exact labels

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{M}(\hat{y}_i - y_i)^2}{M}}$$

The smaller RMSE means the better predictive power

RMSE: Example

- Our true labels

$$y_1 = 3, y_2 = -0.5, y_3 = 2, y_4 = 7$$

- Predictor A gives the predicted labels:

$$\hat{y}_1 = 2.5, \hat{y}_2 = 0, \hat{y}_3 = 2, \hat{y}_4 = 8$$

RMSE for predictor $\mathrm{RMSE_A} = \sqrt{\frac{\sum_{i=1}^{4}(\hat{y}_i - y_i)^2}{4}} = 0.612$

- Predictor B gives the predicted labels:

$$\tilde{y}_1 = 1.5, \tilde{y}_2 = 1.0, \tilde{y}_3 = 2, \tilde{y}_4 = 4$$

RMSE for predictor $\mathrm{RMSE_B} = \sqrt{\frac{\sum_{i=1}^{4}(\tilde{y}_i - y_i)^2}{4}} = 1.837$

Predictor A is better than predictor B.

Pearson Correlation

- Assume we have M true labels

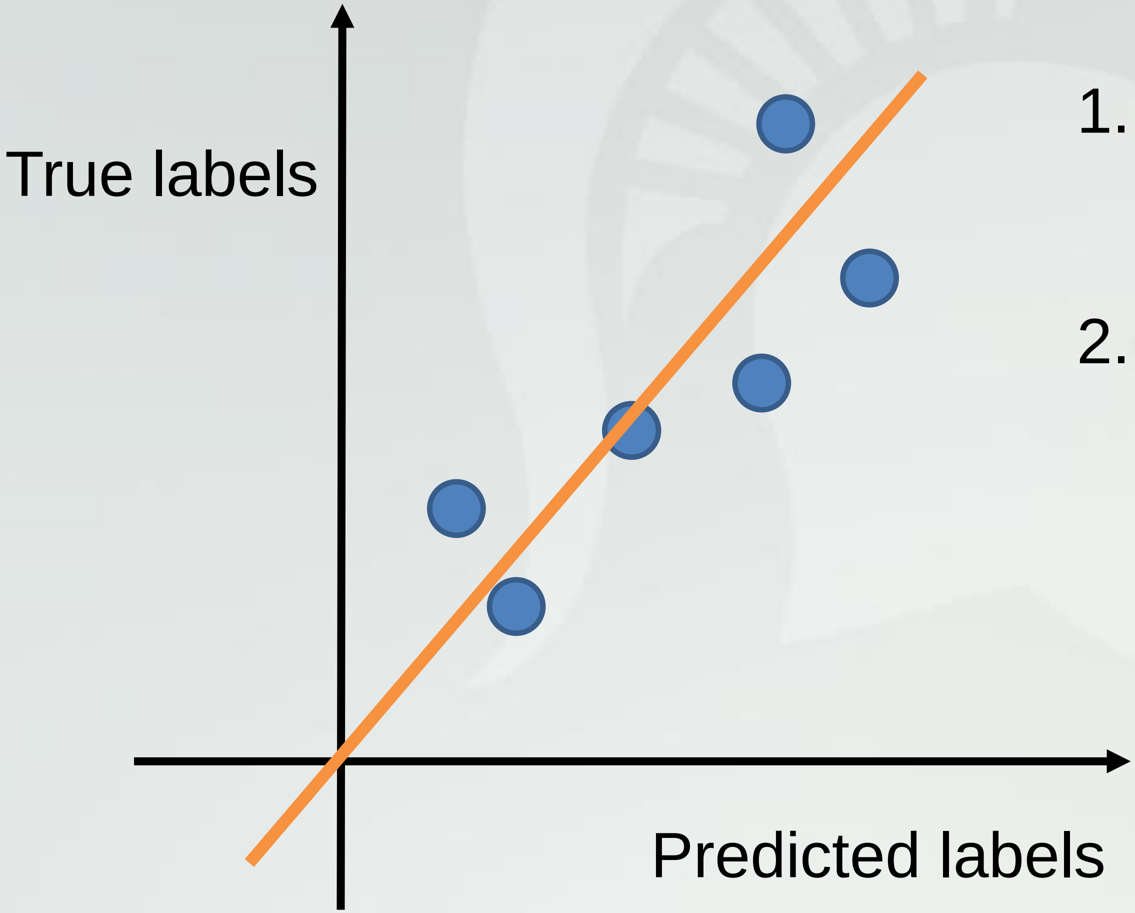
$$y_1, y_2, \ldots, y_M$$

Our predictor gives M predicted labels

$$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_M$$

- Pearson correlation measures the linear correlation between two vectors $(y_1, y_2, \ldots, y_M)$ and $(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_M)$

Interpretation



In Pearson Correlation calculation

1. Draw a best fitting line to the data (how?)

2. Pearson Correlation is the value used to measure how far the data points from the best fitting line

Formulation

- Assume we have M true labels

$$y_1, y_2, \dots, y_M$$

Our predictor gives M predicted labels

$$\hat{y}_1, \hat{y}_2, \dots, \hat{y}_M$$

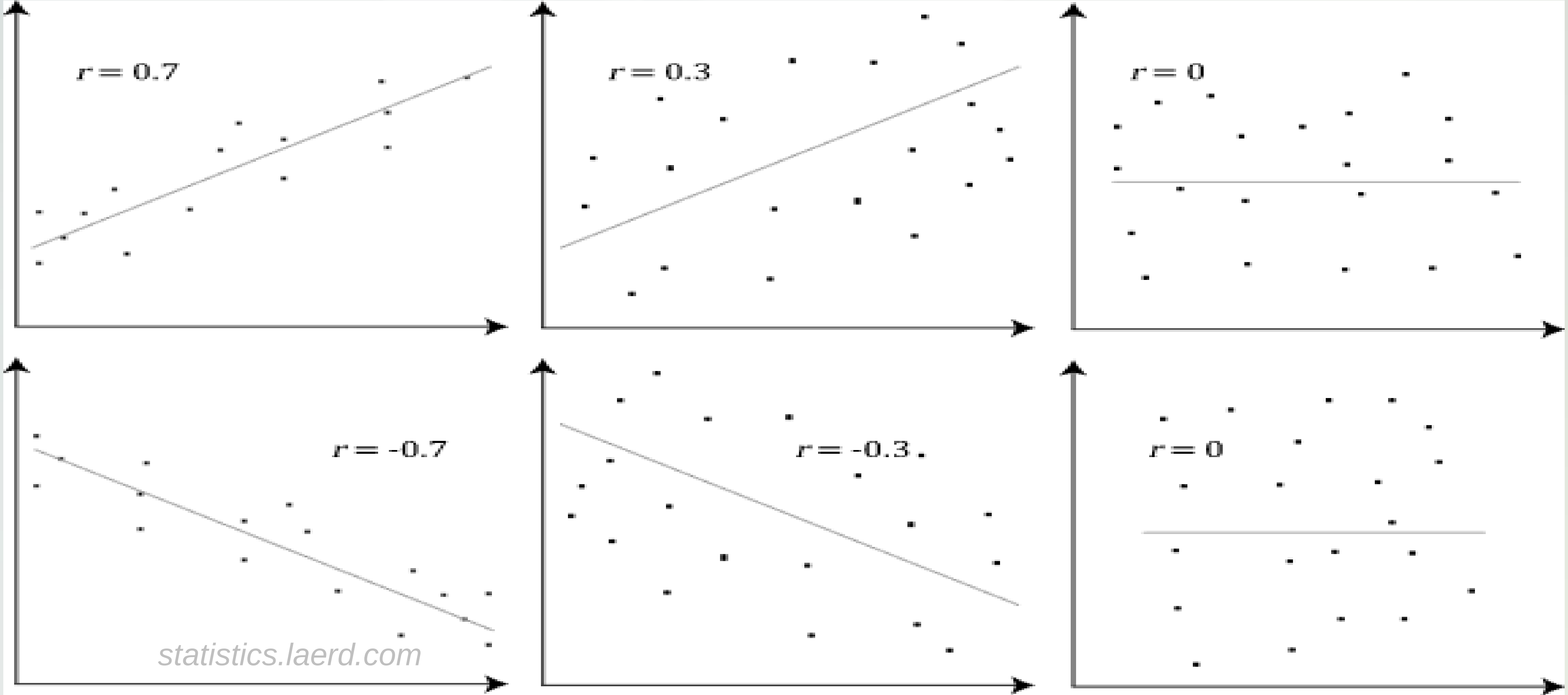
$$R_p = \frac{\sum_{i=1}^{M}(\hat{y}_i - \hat{\mu})(y_i - \mu)}{\sqrt{\sum_{i=1}^{M}(\hat{y}_i - \hat{\mu})^2}\sqrt{\sum_{i=1}^{M}(y_i - \mu)^2}}$$

where

$$\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M}\hat{y}_i, \qquad \mu = \frac{1}{M}\sum_{i=1}^{M}y_i$$

Range

- $-1 \leq R_p \leq 1$
- $R_p > 0$: positive correlation
- $R_p < 0$: negative correlation
- $R_p = 0$: no correlation



statistics.laerd.com

Example

- Our true labels

$$y_1 = 3, y_2 = -0.5, y_3 = 2, y_4 = 7$$

- Predictor A gives the predicted labels

$$\hat{y}_1 = 2.5, \hat{y}_2 = 0, \hat{y}_3 = 2, \hat{y}_4 = 8$$

Pearson correlation of the Predictor A

$$R_p(A) = 0.985$$

- Predictor B gives the predicted labels

$$\tilde{y}_1 = 1.5, \tilde{y}_2 = 1.0, \tilde{y}_3 = 2, \tilde{y}_4 = 4$$

Pearson correlation of the Predictor B

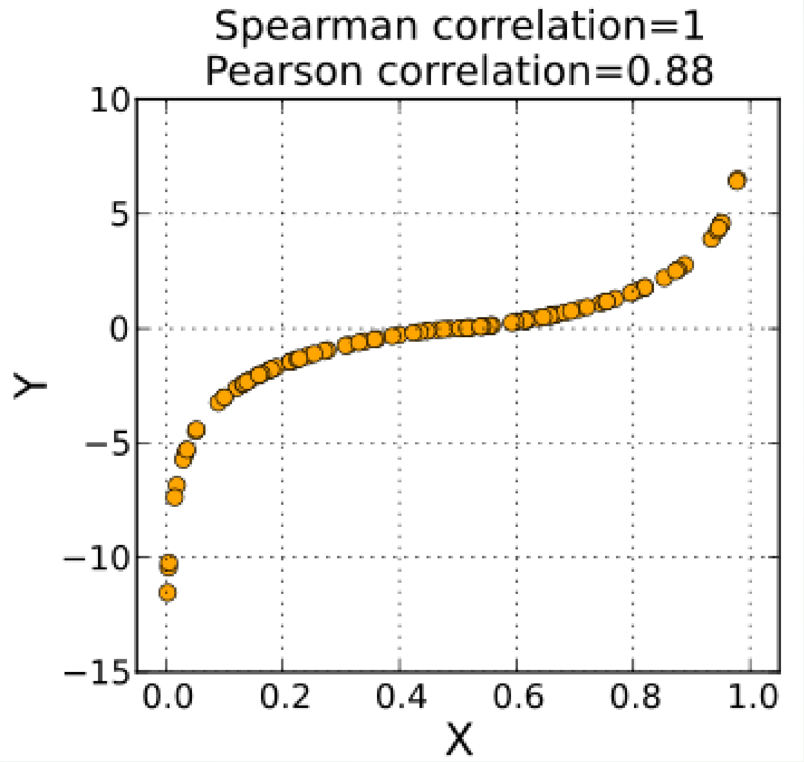
$$R_p(B) = 0.940$$

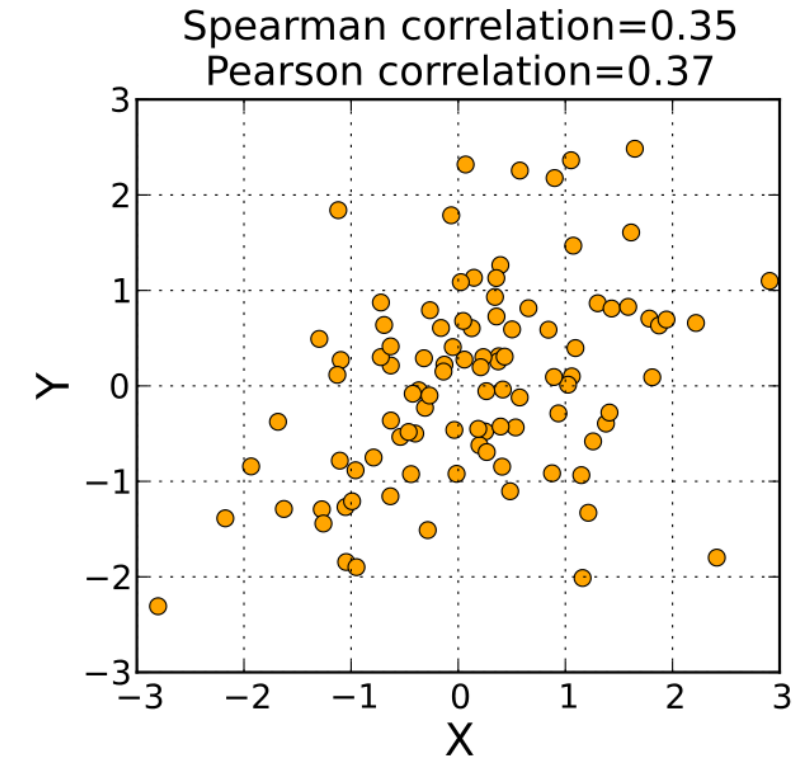
Predictor A is better than predictor B

Spearman Correlation

- Spearman correlation measures monotonic relationship

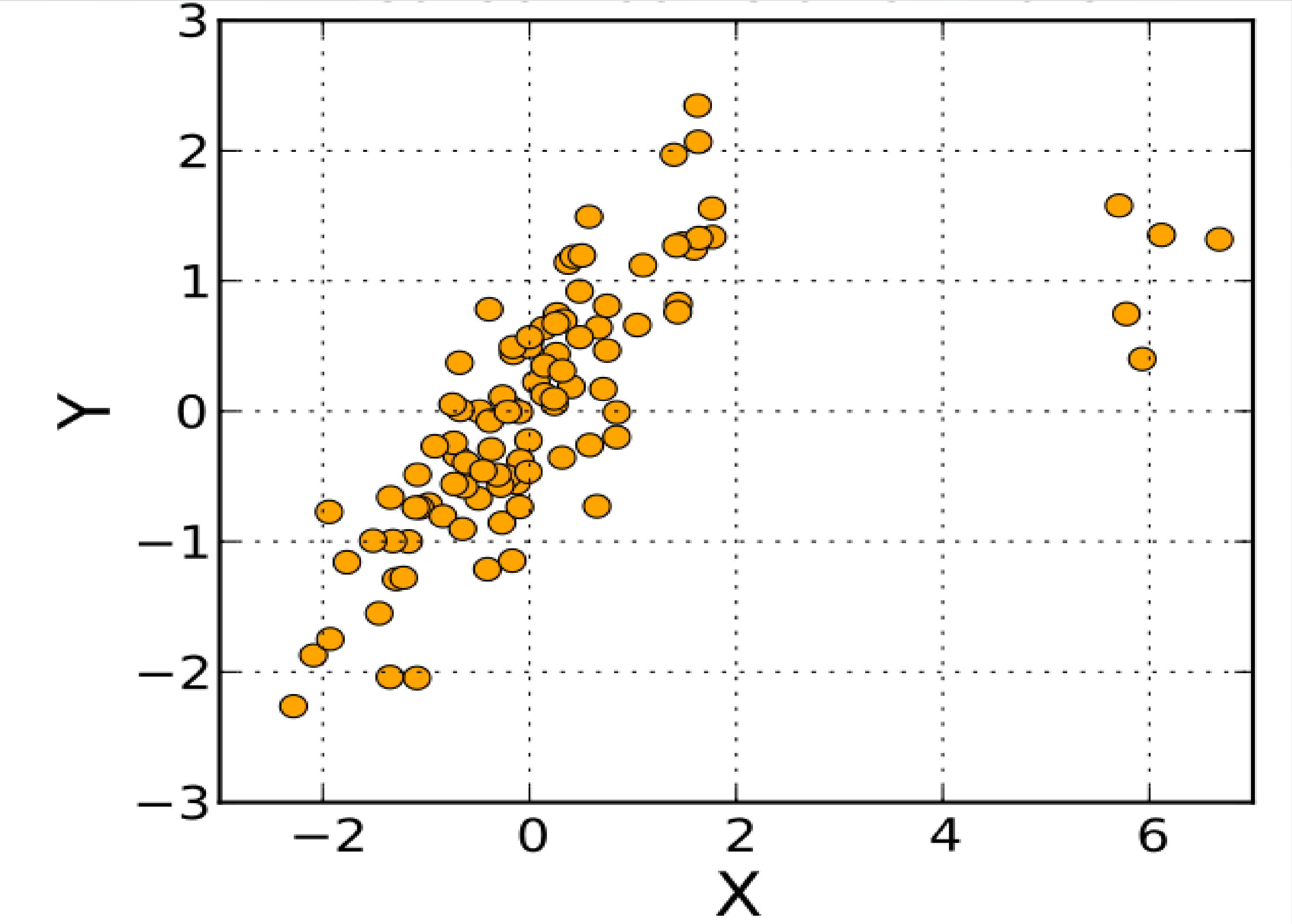
While Pearson correlation measures the linear relationship





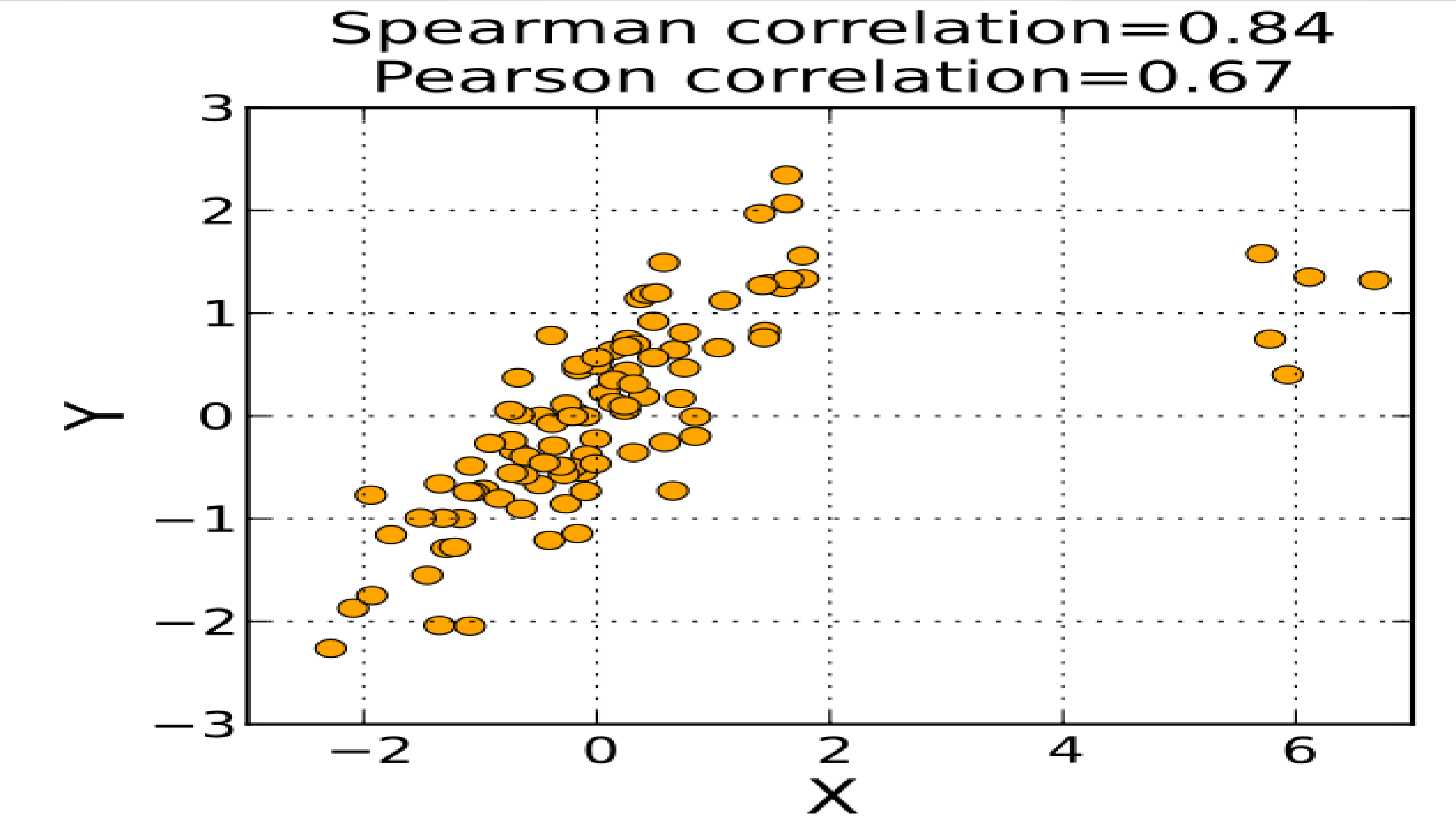
https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Spearman Correlation



https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Spearman Correlation



https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Formulation

- Spearman correlation is considered as the Pearson correlation of the rank values of variables
- Assume we have M true labels

$$y_1, y_2, \dots, y_M$$

Our predictor gives M predicted labels

$$\hat{y}_1, \hat{y}_2, \dots, \hat{y}_M$$

And their ranks are

$r_1, r_2, . . , r_M$ and $\hat{r}_1, \hat{r}_2, \dots, \hat{r}_M$, respectively.

$$R_s = \frac{\sum_{i=1}^{M}(\hat{r}_i - \hat{\mu})(r_i - \mu)}{\sqrt{\sum_{i=1}^{M}(\hat{r}_i - \hat{\mu})^2}\sqrt{\sum_{i=1}^{M}(r_i - \mu)^2}}$$

- Range: $-1 \leq R_s \leq 1$

Example

- Our true labels

$$y_1 = 3, y_2 = -0.5, y_3 = 2, y_4 = 7$$

Predictor A gives the predicted labels

$$\hat{y}_1 = 2.5, \hat{y}_2 = 0, \hat{y}_3 = 2, \hat{y}_4 = 2$$

- Get rank of values

$$y_4 > y_1 > y_3 > y_2$$

$$rank(y_1) = 2, rank(y_2) = 4, rank(y_3) = 3, rank(y_4) = 1$$

$$\hat{y}_1 > \hat{y}_3 = \hat{y}_4 > \hat{y}_2$$

$$rank(\hat{y}_1) = 1, rank(\hat{y}_2) = 4, rank(\hat{y}_3) = 2.5, rank(\hat{y}_4) = 2.5$$

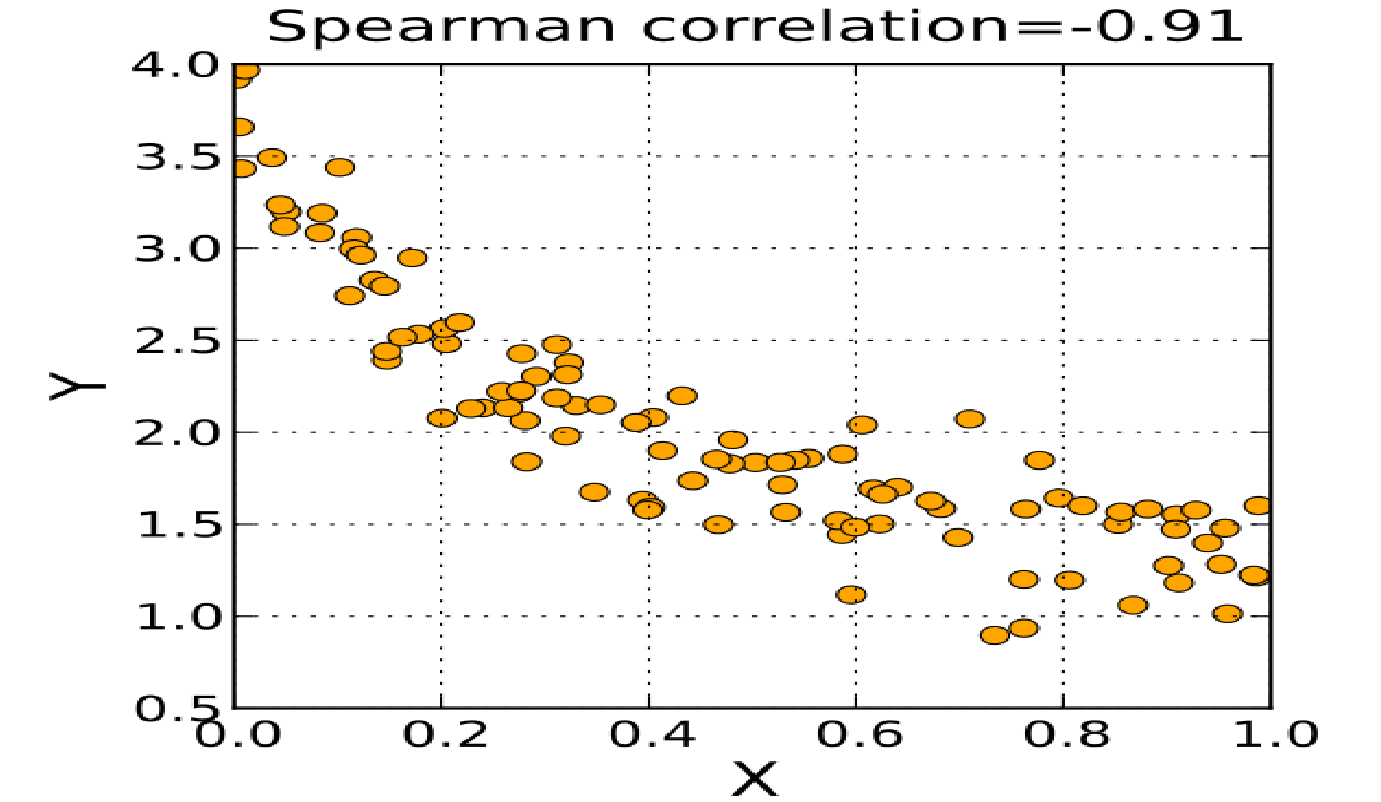
- Spearman correlation = Pearson correlation of

$[2,4,3,1]$ and $[1,4,2.5,2.5]$

$$R_s = 0.632$$

- When all ranks are distinct integers

$$R_s = 1 - \frac{6\sum_{i=1}^{M}(\hat{r}_i - r_i)^2}{M(M^2 - 1)}$$



- If we compare performance of two predictors, the higher is the better

Kendall Tau Correlation

- Kendall Tau is denoted by τ
- It measures relationship based on the rank of variables as in Spearman rank

but Kendall Tau considers the directional agreement instead of the difference

Formulation

- Assume we have M true labels

$$y_1, y_2, \dots, y_M$$

Our predictor gives M predicted labels

$$\hat{y}_1, \hat{y}_2, \dots, \hat{y}_M$$

- Any observation pairs $(y_i, \hat{y}_i)$ and $(y_j, \hat{y}_j)$, $i \neq j$ are said to be
  - concordant: if both $y_i > y_j$ and $\hat{y}_i > \hat{y}_j$ or both $y_i < y_j$ and $\hat{y}_i < \hat{y}_j$
  - discordant: if both $y_i > y_j$ and $\hat{y}_i < \hat{y}_j$ or both $y_i < y_j$ and $\hat{y}_i > \hat{y}_j$
  - neither concordant or discordant: if $y_i = y_j$ or $\hat{y}_i = \hat{y}_j$

Formulation

- $P =$ # of concordant pairs,
- $Q =$ # of discordant pairs
- Kendall Tau τ is defined as

$$\tau = \frac{P - Q}{M(M-1)/2}$$

- Kendall Tau accounting for ties, called Tau-b (τ_b)

$$\tau_b = \frac{P - Q}{\sqrt{P + Q + Y_0}\sqrt{P + Q + \hat{Y}_0}}$$

where Y_0: # of ties only in y variables

$\hat{Y}_0$: # of ties only in $\hat{y}$ variables.

We do not count the ties in both y and $\hat{y}$ variables

Example

Our true labels

$$y_1 = 2, y_2 = -1, y_3 = 1, y_4 = 4$$

Predictor A gives the predicted labels

$$\hat{y}_1 = 1, \hat{y}_2 = 0, \hat{y}_3 = 2, \hat{y}_4 = 2$$

# of concordant pairs = 4

# of discordant pairs = 1;

# of ties in y variables =0;

# of ties in $\hat{y}$ variables =1;

$M = 4$

$$\tau = \frac{4-1}{4 \times 3/2} = 0.5, \tau_b = \frac{4-1}{\sqrt{4+1+0} \times \sqrt{4+1+1}} \sim 0.548$$

Evaluations for Classifiers

- Example

  Our true labels

  $$[0, 0, 0, 0, 1, 1, 1, 1, 1, 1]$$

  Predicted labels

  $$[0, 1, 0, 1, 1, 1, 0, 1, 1, 1]$$

- **Accuracy:**

  Count how many correctly predicted labels

  $$Accuracy = \frac{7}{10} = 0.7$$

Confusion Matrix

- Confusion matrix is a table represents the details about the performance of algorithm on each label

Our true labels

$$[0, 0, 0, 0, 1, 1, 1, 1, 1, 1]$$

Predicted labels

$$[0, 1, 0, 1, 1, 1, 0, 1, 1, 1]$$

N=10	Predicted as 0	Predicted as 1
True label :0	2	2
True label: 1	1	5

True/False Positive/Negative

N=10	Predicted as 0	Predicted as 1
True label :0	2 (True Negative (TN))	2 (False Positive (FP))
True label: 1	1 (False Negative (FN))	5 (True Positive (TP))

N=10	Predicted as 0	Predicted as 1
True label :0	TNR=TN/(TN+FP)	FPR=FP/(TN+FP)
True label: 1	FNR=FN/(FN+TP)	TPR=TP/(FN+TP)

True/False Positive/Negative

N=10	Predicted as 0	Predicted as 1
True label :0	2 (True Negative (TN))	2 (False Positive (FP))
True label: 1	1 (False Negative (FN))	5 (True Positive (TP))

N=10	Predicted as 0	Predicted as 1
True label :0	TNR=0.5	FPR=0.5
True label: 1	FNR=0.167	TPR=0.833

Due to the meaningful of each label, we may wish to reduce **FPR** or **FNR**

Receiver Operating Characteristic (ROC) Curve

- In logistic regression, we choose threshold $z = 0.5$

$$p_{\mathbf{c}}(\mathbf{x}) \geq 0.5\text{: label of } \mathbf{x} \text{ is } 1$$
$$p_{\mathbf{c}}(\mathbf{x}) < 0.5\text{: label of } \mathbf{x} \text{ is } 0$$

- If we increase value of threshold z
  - TPR? FPR? TNR? FNR?
- If we decrease value of threshold z
  - TPR? FPR? TNR? FNR?
- When we vary $0 \leq z \leq 1$, we get different pairs (TPR, FPR).

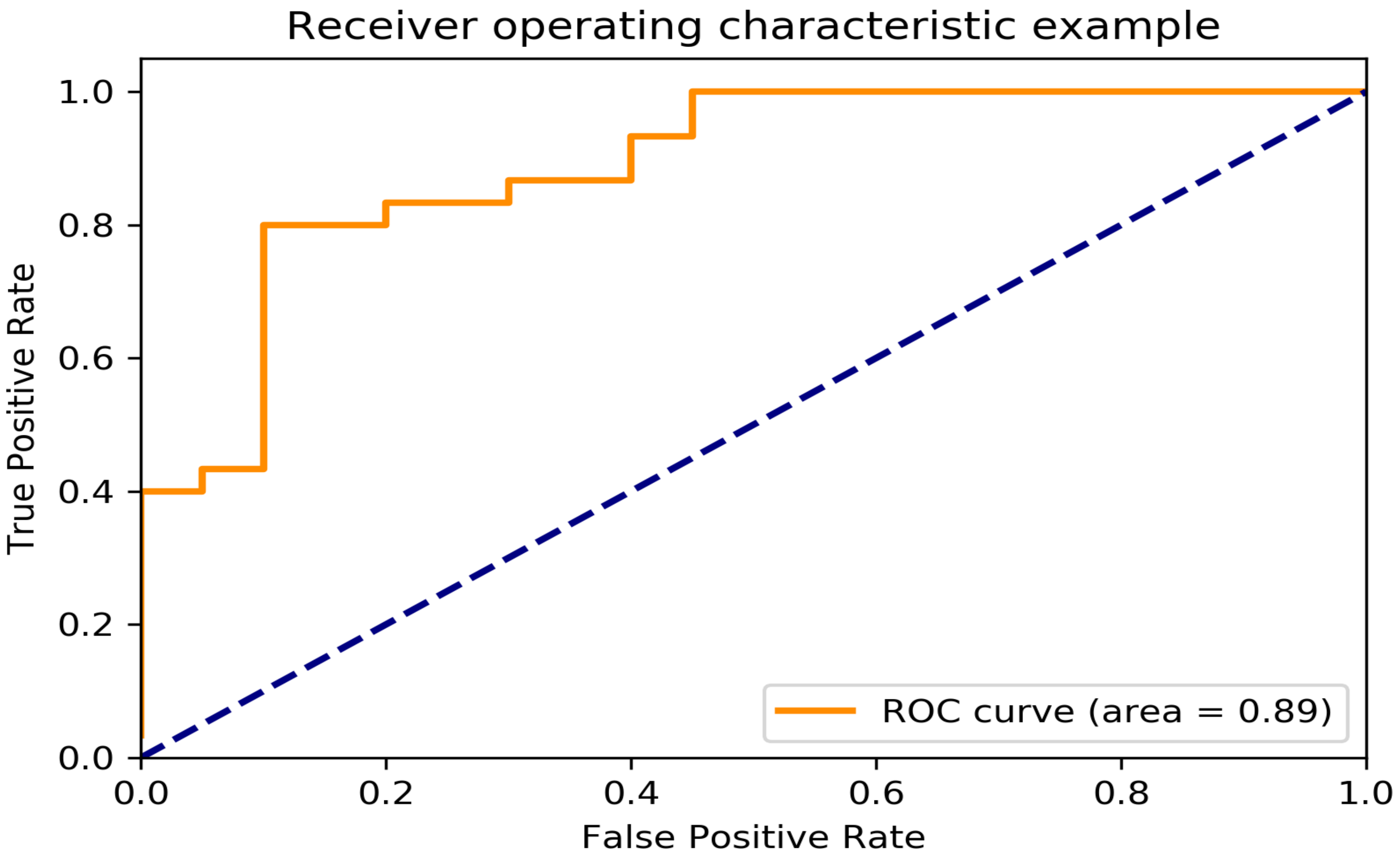
The plot of (TPR, FPR) gives us ROC curve.

ROC and AUC

Threshold	FPR	TPR
0.693	0.0	0.033
0.493	0.0	0.4
0.482	0.05	0.4
….	…	…
0.311	0.3	0.833
…	…	…
0.024	1.0	1.0

ROC and AUC

- AUC = area under the curve



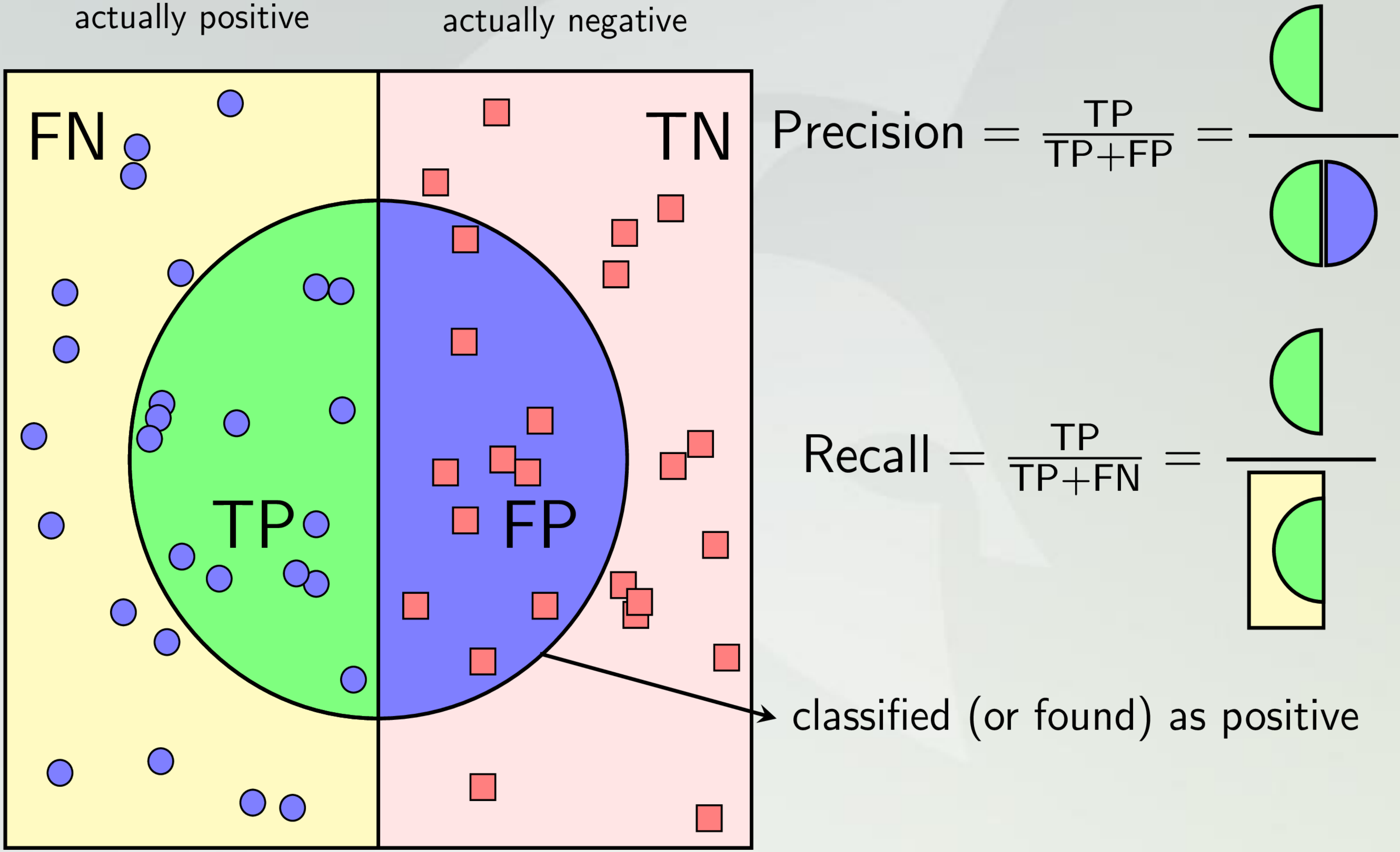
Precision and Recall

- Used when we want to evaluate the performance of predictor on a specific label
- Assume our data has 12 dogs and some cats
- Our predictor identifies 8 dogs, however among these 8 dogs, only 5 ones are true dogs and the rest (3) is cat.

Precision of our predictor= $\frac{5}{8}$

Recall (sensitivity) of our predictor= $\frac{5}{12}$

Precision and Recall



$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

https://machinelearningcoban.com/2017/08/31/evaluation/

Precision and Recall

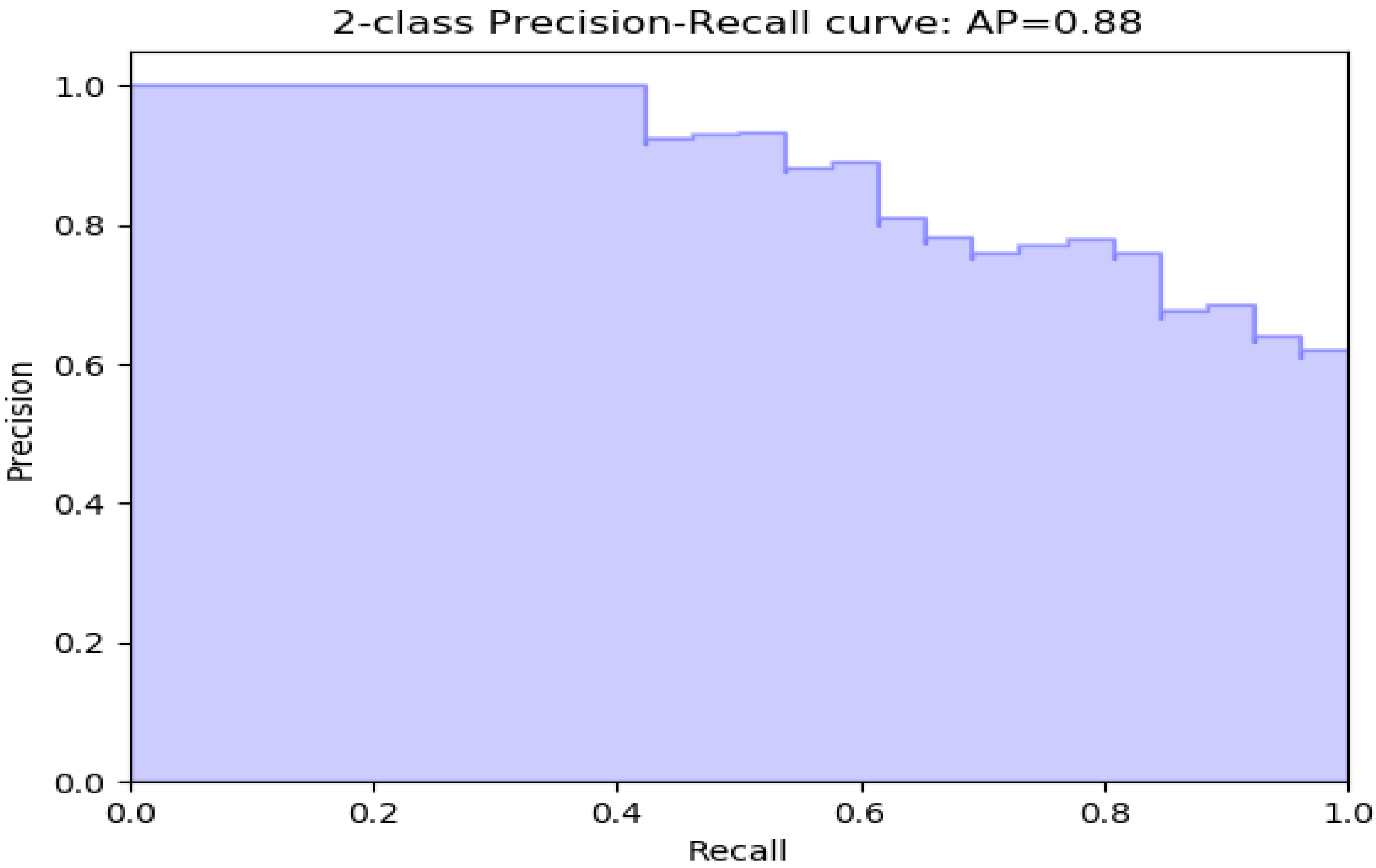
- $0 \leq \text{Precision} \leq 1, 0 \leq \text{Recall} \leq 1$
- $\text{Precision} = 1$ or $\text{Recall} = 1$ does not indicate we have a good predictor
- Precision or Recall does not help us compare the performance between different predictors
- A good predictor needs to achieve high values for both Precision and Recall

Precision-Recall Curve and Average Precision (AP)

- For a predefined threshold we have a pair value (Precision, Recall).
- Plot of all pairs (Precision, Recall) = Precision-Recall curve
- AUC of Precision-Recall curve = Average Precision (*AP*)
- Consider N different thresholds, we obtain N pairs $(P_i, R_i), 1 \leq i \leq N$

$$AP = \sum_{i=1}^{N-1} (R_{i+1} - R_i) P_i$$

Precision-Recall Curve and Average Precision (*AP*)



F_1 Score

- F_1 score measures the harmonic mean of precision and recall

$$\frac{2}{F_1} = \frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}$$

or

$$F_1 = 2\,\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- $0 \leq F_1 \leq 1$. Higher F_1 value, better predictor

F_1 Score

Precision	Recall	F_1
1	1	1
0.1	0.1	0.1
0.5	0.5	0.5
1	0.1	0.182
0.3	0.8	0.36

https://machinelearningcoban.com/2017/08/31/evaluation/

F_β Score

- A generalized form of F_1 is F_β

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

- When $\beta = 1$ we get F_1 score
- When $\beta > 1$ we emphasize precision
- When $\beta < 1$ we emphasize recall

How to Evaluate Predictor on Multi-labels datasets

- Example: our true labels

$$[1,0,1,1,2,2,1,2,0,1]$$

Our predicted labels

$$[1,1,2,1,0,2,1,2,0,2]$$

N=10	Predicted as 0	Predicted as 1	Predicted as 2
True label :0	1	1	0
True label: 1	0	3	2
True label: 2	1	0	2

$$\text{Accuracy} = \frac{1+3+2}{10} = 0.6$$

Micro-Average Precision and Micro-Average Recall

- $\mathrm{Total\ TP} = \mathrm{TP}^0 + \mathrm{TP}^1 + \mathrm{TP}^2$
- $\mathrm{Total\ FP} = \mathrm{FP}^0 + \mathrm{FP}^1 + \mathrm{FP}^2$
- $\mathrm{Total\ FN} = \mathrm{FN}^0 + \mathrm{FN}^1 + \mathrm{FN}^2$

$$\mathrm{Micro} - \mathrm{average\ Precision} = \frac{\mathrm{Total\ TP}}{\mathrm{Total\ TP} + \mathrm{Total\ FP}}$$

$$\mathrm{Micro} - \mathrm{average\ Recall} = \frac{\mathrm{Total\ TP}}{\mathrm{Total\ TP} + \mathrm{Total\ FN}}$$

Macro-Average Precision and Macro-Average Recall

$$\mathrm{Macro} - \mathrm{Average\ Precision} = \frac{\mathrm{P}^0 + \mathrm{P}^1 + \mathrm{P}^2}{3}$$

$$\mathrm{Macro} - \mathrm{Average\ Recall} = \frac{\mathrm{R}^0 + \mathrm{R}^1 + \mathrm{R}^2}{3}$$

Where P^i and R^i are the precision and recall for each label, respectively.