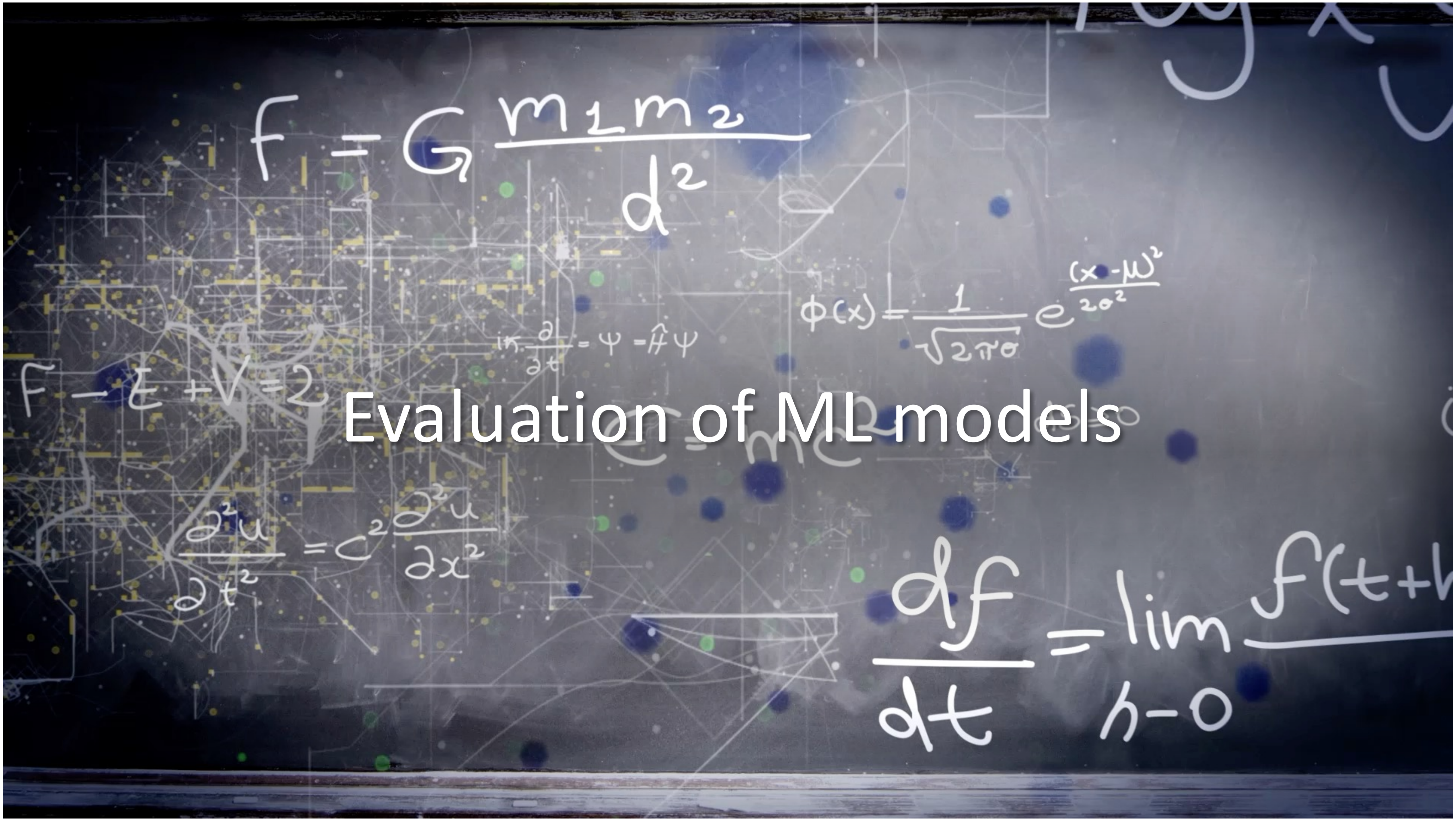


Evaluation of ML models

Evaluations of ML models

1. Metrics
2. Confusion Matrix
3. Precision and recall
4. Receiver Operating Characteristic (ROC) curves
5. Bias VS Variance
6. Measuring error in Regression Problems

Metrics

It is extremely important to use **quantitative metrics** for evaluating a machine learning model

Until now, we relied on the **cost function value** for regression and classification

Other metrics can be used to **better evaluate** and understand the model

For classification
✓ Accuracy/Precision/Recall/F1-score, ROC curves,...

For regression
✓ Normalized RMSE, Normalized Mean Absolute Error (NMAE),...

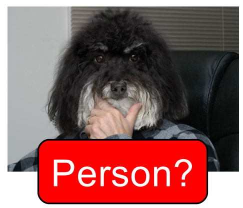
How do we evaluate an ML model?

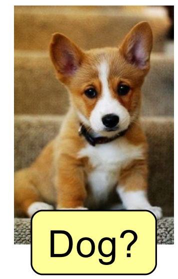
- We measure the accuracy / error on test data:

- Accuracy: 4 correct out of 6 = 66.67%
- Error: 2 wrong out of 6 = 33.33%

How do we evaluate an ML model?

- We build the confusion matrix





- Accuracy: the sum of the elements from the main diagonal divided by the sum of non-zero components (4/6)
- Error: the sum of the elements outside the main diagonal divided by the sum of non-zero components (2/6)

Predicted Actual	Car	Dog	Person
Car	1	1	0
Dog	0	1	1
Person	0	0	2

How do we evaluate an ML model?

- Confusion matrix in the binary case

	Predicted YES	Predicted NO
Actual YES	True Positive	False Negative
Actual NO	False Positive	True Negative

How do we evaluate an ML model?

- Confusion matrix in the binary case

	Predicted YES	Predicted NO
Actual YES	2	False Negative
Actual NO	False Positive	True Negative

How do we evaluate an ML model?

- Confusion matrix in the binary case

	Predicted YES	Predicted NO
Actual YES	2	1
Actual NO	False Positive	True Negative

How do we evaluate an ML model?

- Confusion matrix in the binary case

	Predicted YES	Predicted NO
Actual YES	2	1
Actual NO	1	True Negative

How do we evaluate an ML model?

- Confusion matrix in the binary case

	Predicted YES	Predicted NO
Actual YES	2	1
Actual NO	1	2

How do we evaluate an ML model?

- We compute Precision and Recall

- Precision = TP / (TP + FP)

  = 66.67%

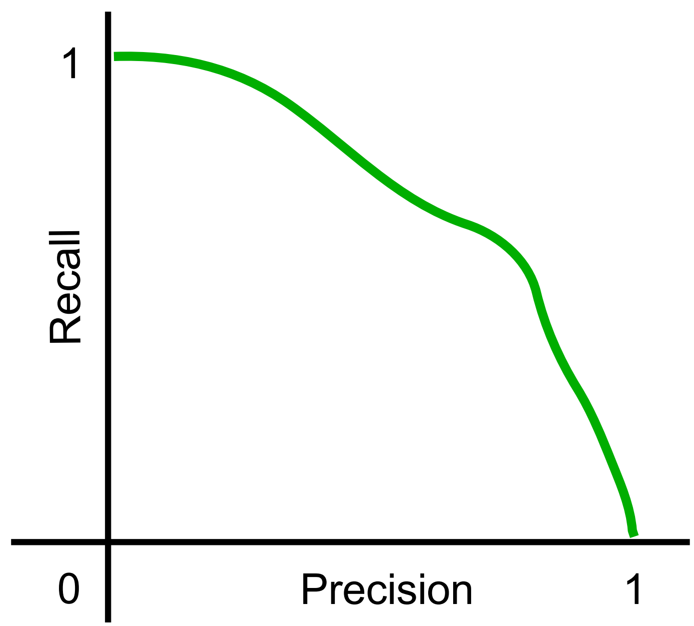
- Recall = TP / (TP + FN)

  = 66.67%

	Predicted YES	Predicted NO
Actual YES	2	1
Actual NO	1	2

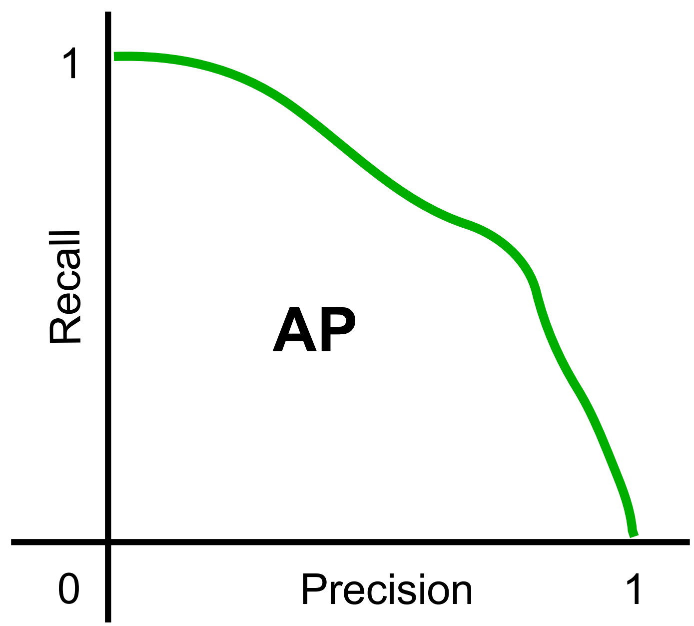
How do we evaluate an ML model?

- Precision-Recall curve



How do we evaluate an ML model?

- Average Precision



How do we evaluate an ML model?

- We compute the TPR and FPR

- TPR = TP / (TP + FP)

  = 66.67%

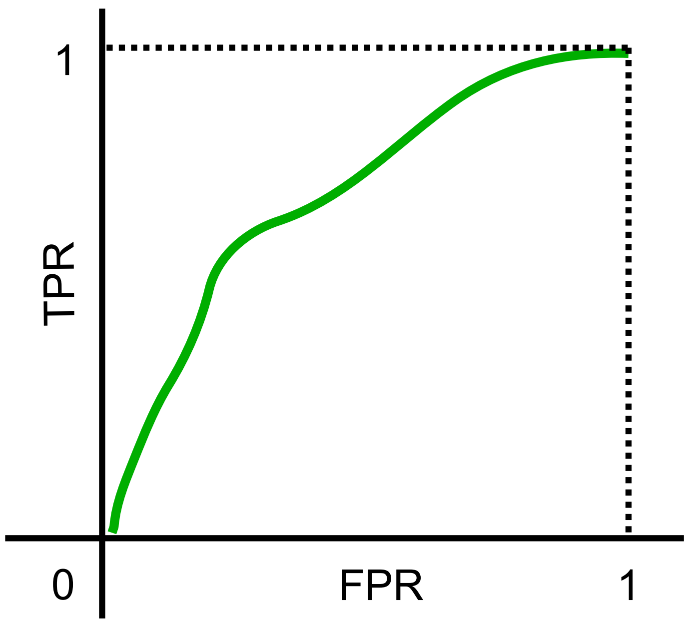
- FPR = FP / (FP + TN)

  = 33.33%

	Predicted YES	Predicted NO
Actual YES	2	1
Actual NO	1	2

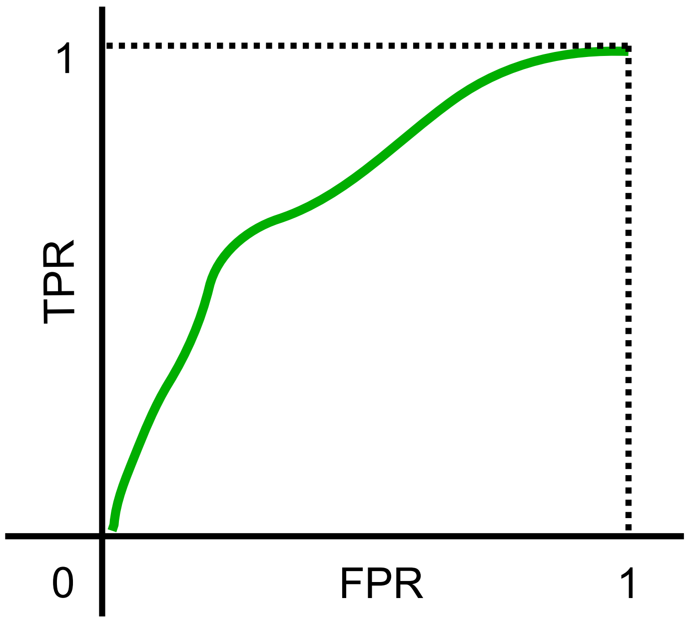
How do we evaluate an ML model?

- ROC (Receiver Operating Characteristic) curve



How do we evaluate an ML model?

- We compute the AUC (area under the ROC curve)



How do we evaluate an ML model?

- We compute the F_β score:

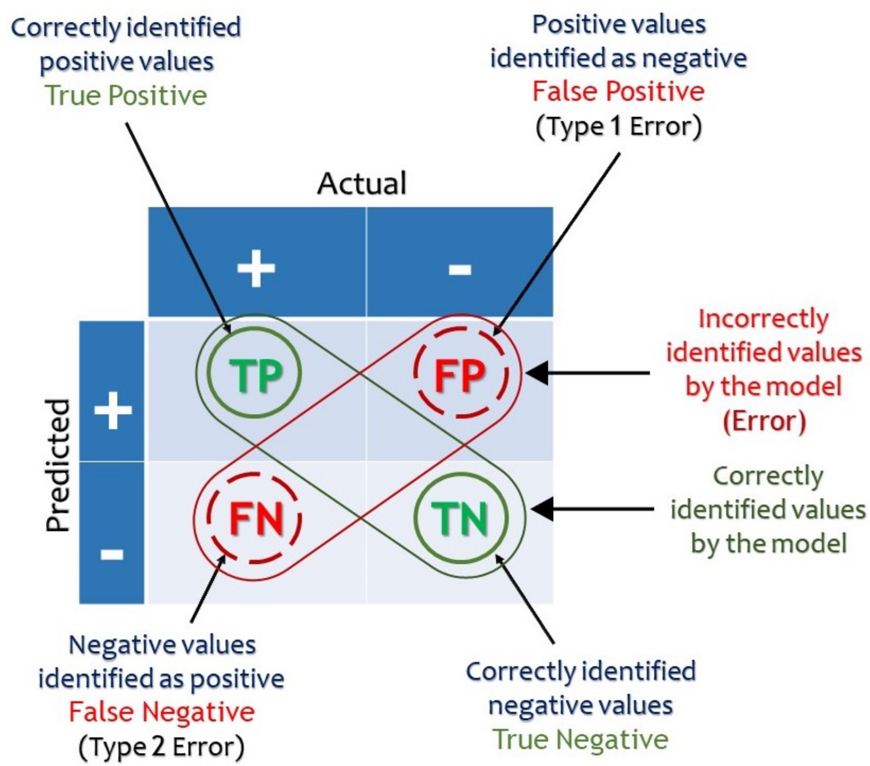
$$F_\beta = (1 + \beta^2) \cdot \frac{\textbf{precision} \cdot \textbf{recall}}{(\beta^2 \cdot \textbf{precision}) + \textbf{recall}}$$

- When $\beta < 1$, precision is more important
- The F_1 score is the most commonly-used in practice:
  - gives equal importance to precision and recall

Confusion Matrix

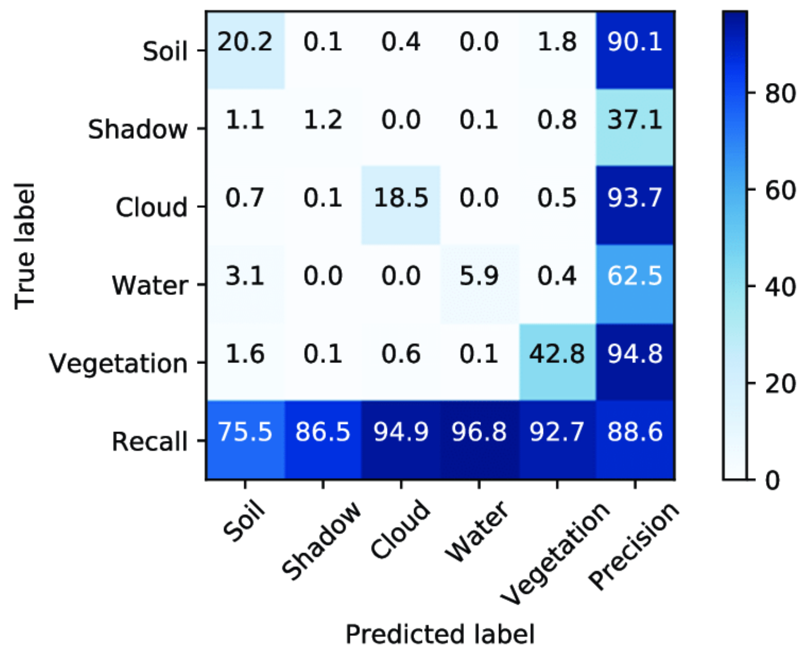
		Actual	
		Dog	Cat
Predicted	Dog	**8**	**2**
	Cat	**1**	**9**

	Actual Positive	Actual Negative
Predicted Positive	True Positive TP	False Positive FP (Type I Error)
Predicted Negative	False Negative FN (Type II Error)	True Negative TN


Correctly identified
positive values
True Positive
Positive values
identified as negative
False Positive
(Type 1 Error)
Actual
+
-
Predicted
+
-
TP
FP
FN
TN
Incorrectly
identified values
by the model
(Error)
Correctly
identified values
by the model
Negative values
identified as positive
False Negative
(Type 2 Error)
Correctly identified
negative values
True Negative

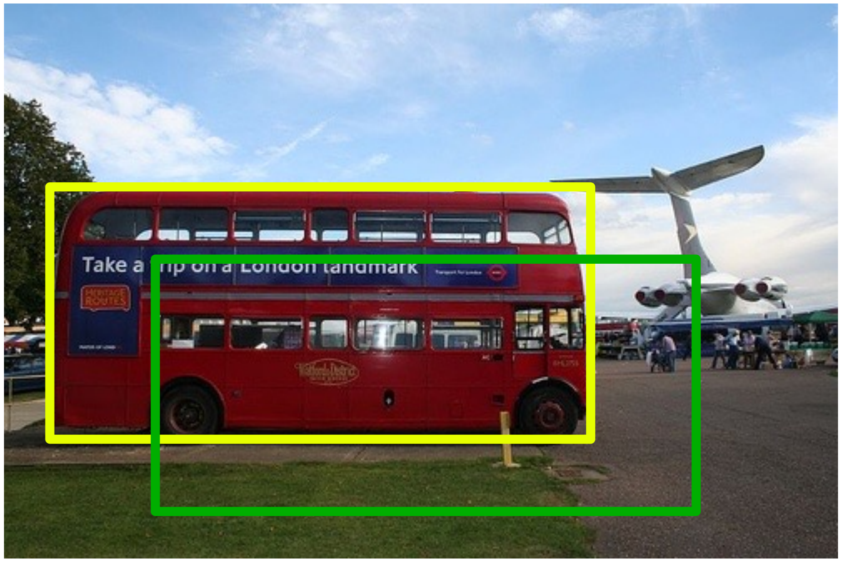
Predicted Class

		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	**Negative Predictive Value** $\frac{TN}{(TN + FN)}$	**Accuracy** $\frac{TP + TN}{(TP + TN + FP + FN)}$


Soil
Shadow
Cloud
Water
Vegetation
Recall
True label
Predicted label
Precision
20.2 0.1 0.4 0.0 1.8 90.1
1.1 1.2 0.0 0.1 0.8 37.1
0.7 0.1 18.5 0.0 0.5 93.7
3.1 0.0 0.0 5.9 0.4 62.5
1.6 0.1 0.6 0.1 42.8 94.8
75.5 86.5 94.9 96.8 92.7 88.6
80
60
40
20
0

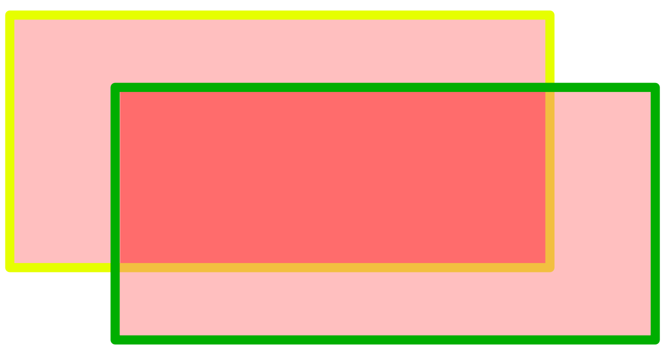
How do we evaluate a detection model?

- Intersection over Union (Jaccard index)

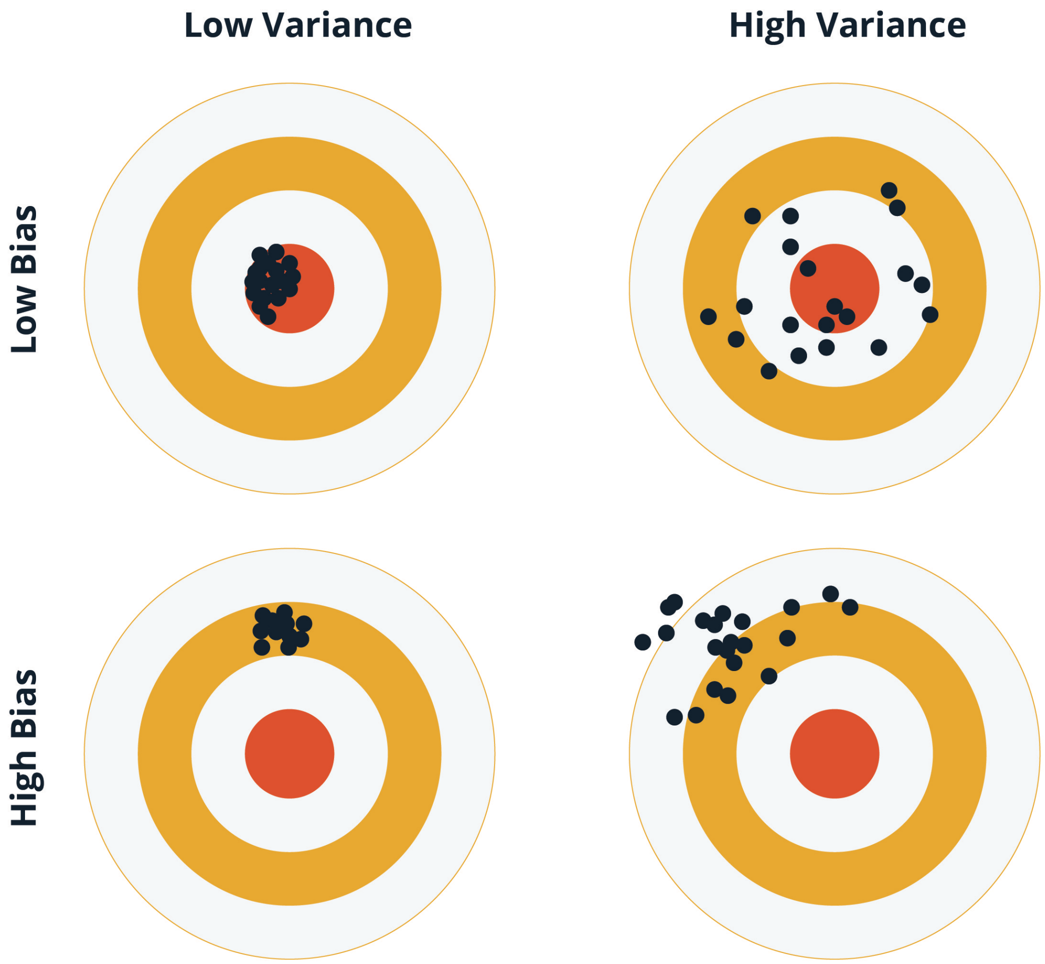


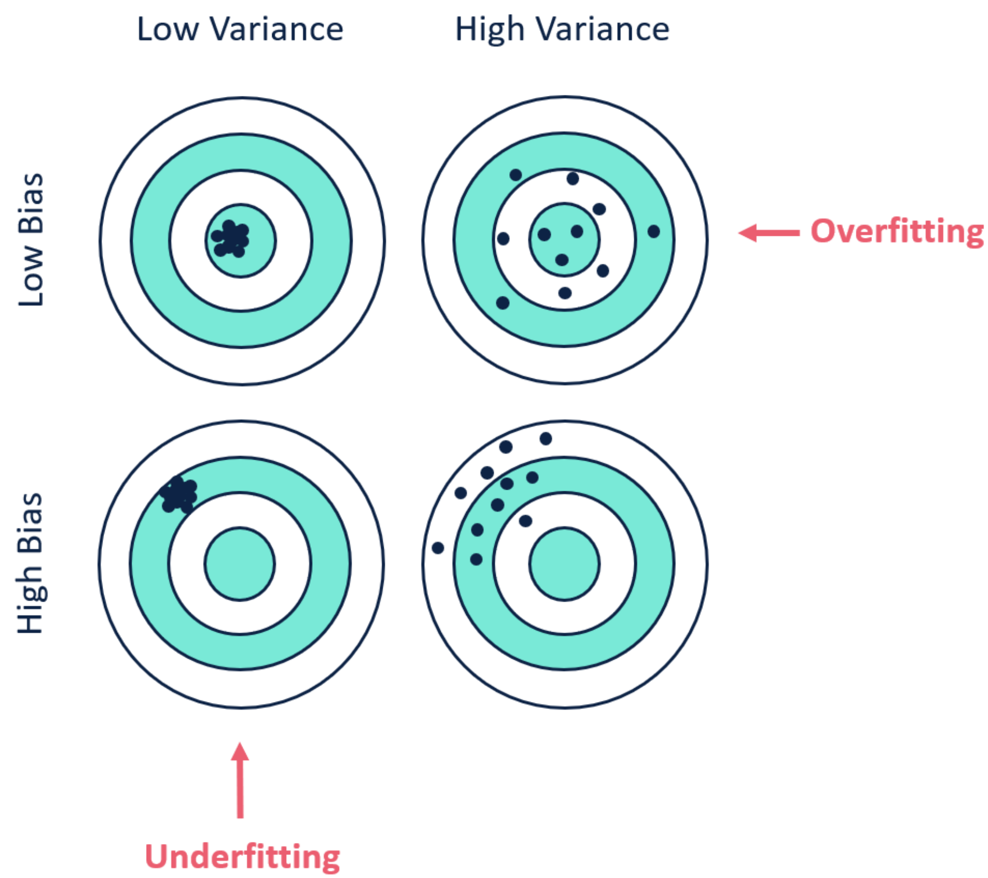
How do we evaluate a detection model?

- Intersection over Union (Jaccard index)
- Correct detection if J(A,B) > 0.5

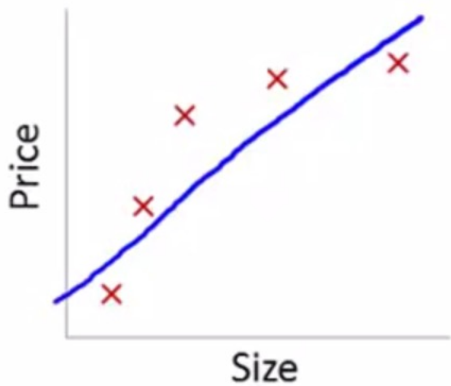


Bias vs Variance


Low Variance
High Variance
Low Bias
High Bias


Low Variance
High Variance
Low Bias
High Bias
Overfitting
Underfitting

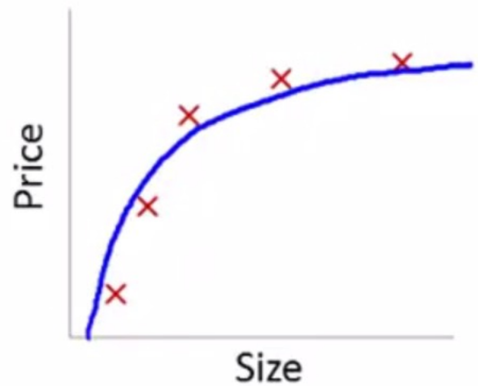
Bias/variance



$$\theta_0 + \theta_1 x$$

High bias
(underfit)

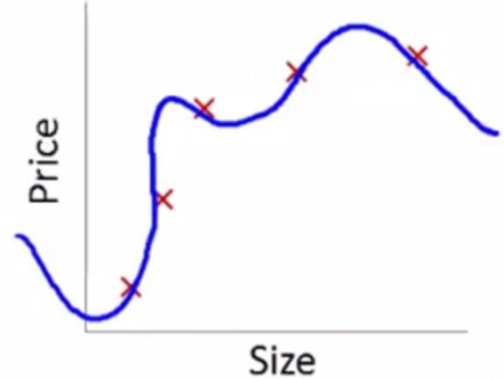
$d=1$



$$\theta_0 + \theta_1 x + \theta_2 x^2$$

"Just right"

$d=2$



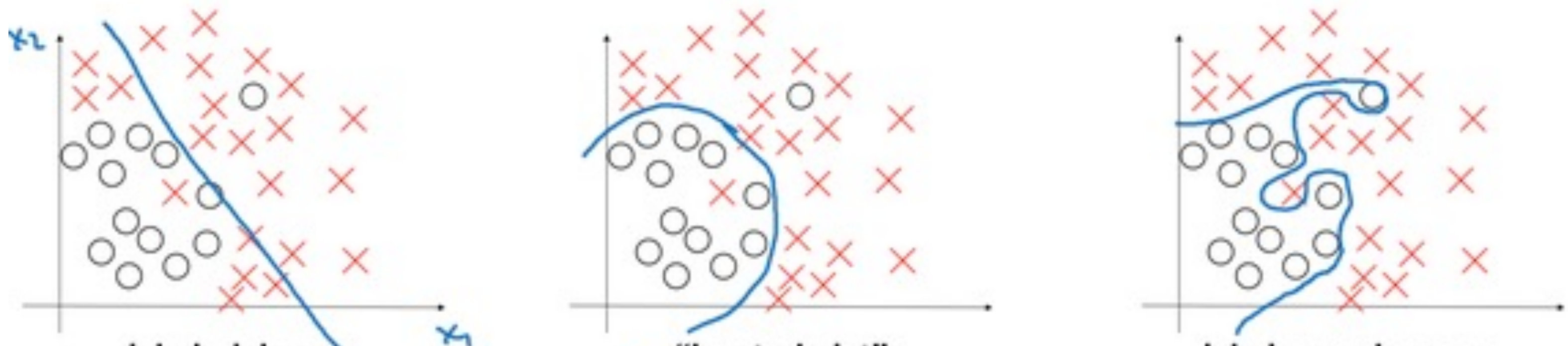
$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

High variance
(overfit)

$d=4$

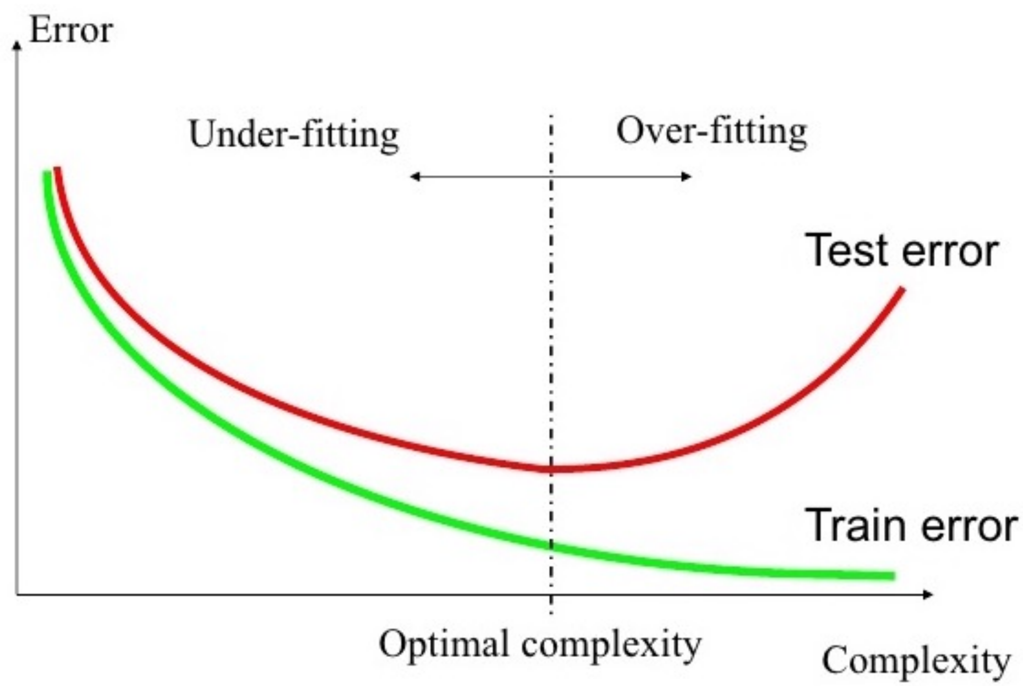
Andrew Ng

Bias and Variance



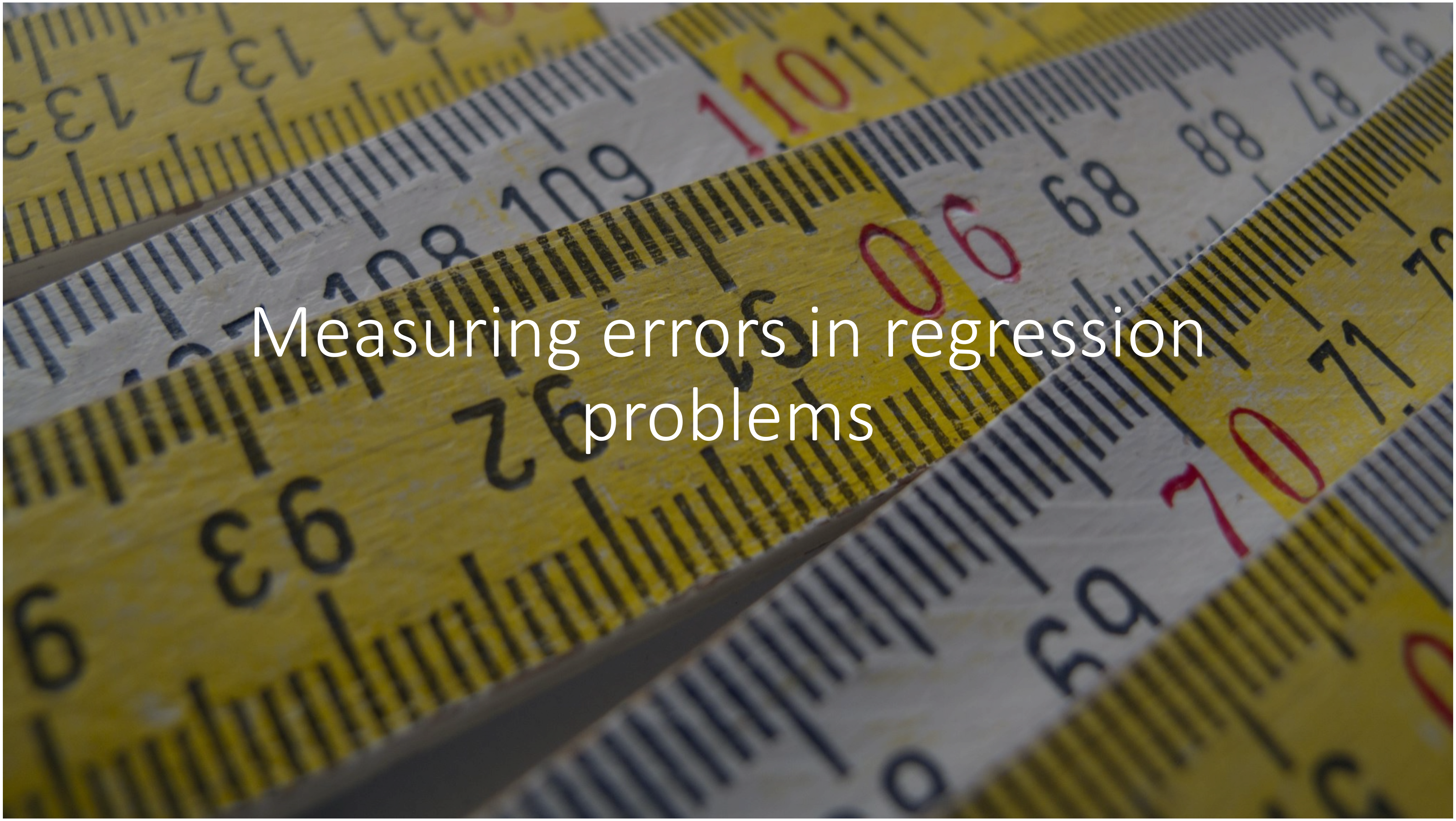
high bias | "just right" | high variance

underfitting | | overfitting



Error
Under-fitting
Over-fitting
Test error
Train error
Optimal complexity
Complexity

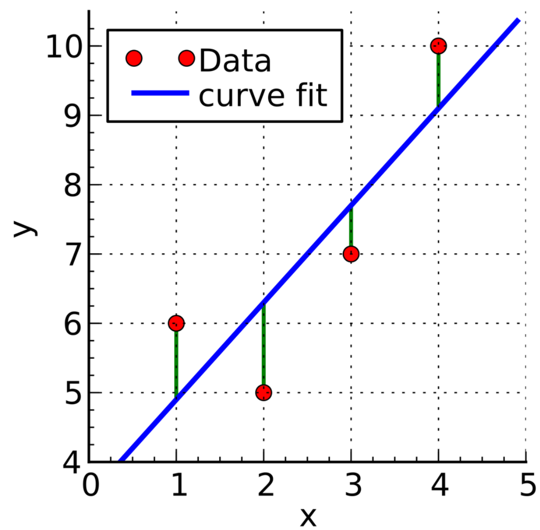
Measuring errors in regression problems



How do we evaluate a regression model?

- Mean squared error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$



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

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

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

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

Where,

$\hat{y}$ $-$ *predicted value of y*

$\bar{y}$ $-$ *mean value of y*

Summary

1. Metrics in ML
2. Confusion Matrix
3. Precision and recall
4. Receiver Operating Characteristic (ROC) curves
5. Bias VS Variance
6. Measuring error in Regression Problems