Amirkabir University of Technology
(Tehran Polytechnic)

Machine Learning

Lecture 7.
Supervised learning
Evaluation

Alireza Rezvanian

Fall 2023

Amirkabir University of Technology (Tehran Polytechnic)

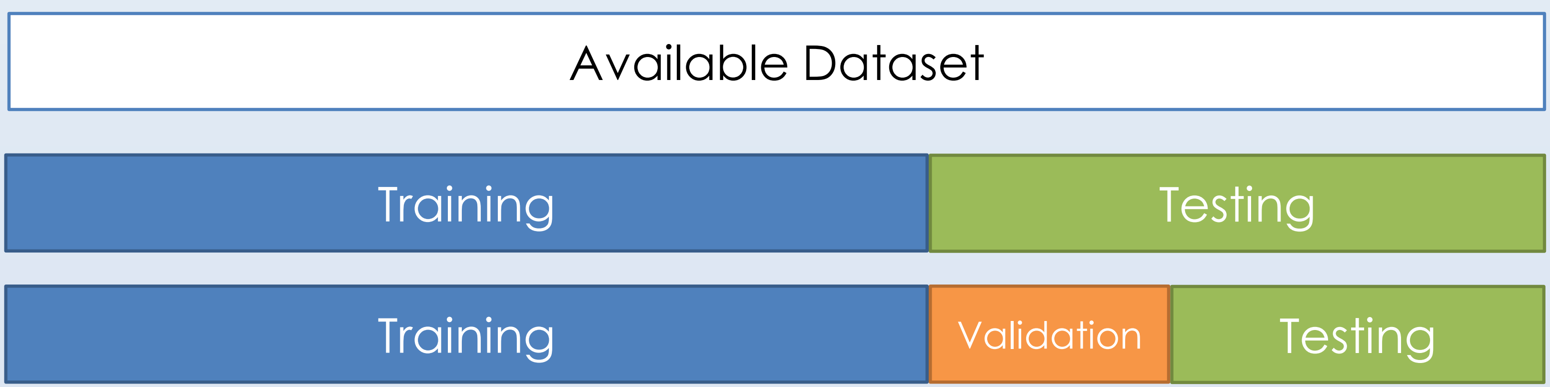
Last update: Oct. 22, 2021

Outline

- Hold-out method
- K-fold cross validation
- Accuracy
- Error
- Precision
- Recall
- F-measure

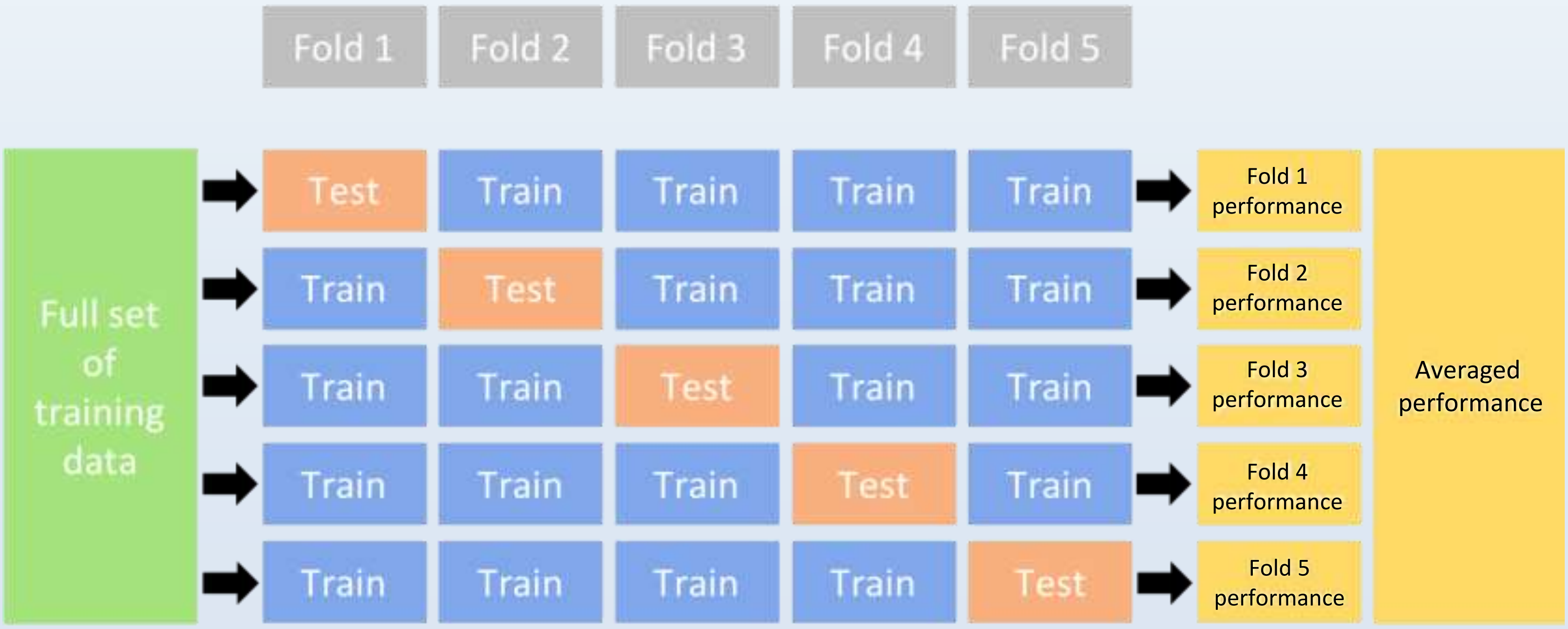
Evaluation

- To compare different models.
- To tune the hyper-parameters such as
  - K in KNN, number of layers in neural networks, the best pruning of a decision tree, etc.
- The main goal of ML is generalization. We want to measure the generalization ability of our model.
- **Hold-out** method: You train on the Training data and evaluate your model on the Testing data. Once your model is ready, you test it one final time on the test data.
- Shuffle data before splitting



K-fold cross validation

- When you have few data points, the validation set would end up being very small. This would prevent you from reliably evaluating your model. So, we use k-fold cross validation.
- Typical values for k: 5, 10, N (leave-one-out method



Evaluation metrics

Confusion Matrix

	Real Positive (1)	Real Negative (0)
Predicted Positive (1)	True Positive (TP)	False Positive (FP)
Predicted Negative (0)	False Negative (FN)	True Negative (TN)
	TP + FN = P	FP + TN = N

P: the number of real positive cases in the data
N: the number of real negative cases in the data

TP: True Positive
TN: True Negative
FP: False Positive (Type **I** error)
FN: False Negative (Type **II** error)

Accuracy

- Percentage of instances that are correctly classified

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

$$Error = 1 - Accuracy = \frac{FP + FN}{TP + FP + TN + FN}$$

TP rate = $TPR = \frac{TP}{P} = \frac{TP}{TP+FN} = Sensitivity$

TN rate = $TNR = \frac{TN}{N} = \frac{TN}{TN+FP} = Specificity$

- Is Accuracy (Error) always a good measure?
  - Consider a cancer detection system always predicts "no cancer"
  - Not a good measure for imbalanced data!

Precision

- Percentage of instances that the classifier labeled as positive are actually positive

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

	Actually Spam = (Yes)	**Actually Spam = (No)**	**Total**
Predicted Spam = (yes)	**60 (TP)**	**140 (FP)**	**200**
Predicted Spam = (No)	120 (FN)	680 (TN)	800
Total	180	820	1000

$$Precision = \frac{TP}{TP + FP} = \frac{60}{60 + 140} = 0.3$$

Recall

- Percentage of positive instances that the classifier labeled as positive are actually positive

$$Recall = \frac{TP}{TP + FN} = TPR = Sensitivity$$

	Actually Spam = (Yes)	Actually Spam = (No)	Total
Predicted Spam = (yes)	**60 (TP)**	140 (FP)	200
Predicted Spam = (No)	**120 (FN)**	680 (TN)	800
Total	**180**	820	1000

$$Recall = \frac{TP}{TP + FN} = \frac{60}{60 + 120} = 0.33$$

F-measure

- Is it enough to have good precision or good recall?
- We should combine precision and recall into one measure.
- The most popular way is by harmonic mean: F-measure

$$F\text{–}measure = \frac{2 \times Precision \times Recall}{Precision + Recall} = F\text{–}Score = F_1$$

$$F\text{–}measure = \frac{2\times 0.3\times 0.33}{0.3+0.33} = 0.31$$

Evaluation in multi class

- Compute all TP, FN and FP as one vs. rest
  - **Micro**
    - Compute cumulative for TP, FN, FP and F-measure
  - **Macro**
    - Take average on each measure
  - **Weighted**
    - Weighted average of each measure for different classes, where the weight of each class is proportional to the number of instances in that class

Evaluation example for multi class

Predicted class = {0, 2, 1, 0, 0, 2, 0}

Actual class = {0, 1, 2, 0, 1, 2, 0}

	Actually C = 0	Actually C =1	Actually C =2
Predicted C = 0	3	1	0
Predicted C = 1	0	0	1
Predicted C = 2	0	1	1

C_0={TP=3, FP=1, FN=0}

$$P_0 = \frac{3}{4} \qquad R_0 = \frac{3}{3} \qquad F_1 0 = 0.86$$

C_1={TP=0, FP=1, FN=2}

$$P_1 = \frac{0}{1} = 0 \quad R_1 = \frac{0}{2} = 0 \qquad F_1 1 = 0$$

C_2={TP=1, FP=1, FN=1}

$$P_2 = \frac{1}{2} \qquad R_2 = \frac{1}{2} \qquad F_1 2 = 0.5$$

$$\text{Micro: } P = \frac{3+0+1}{7} = \frac{4}{7},\ \ R = \frac{3+0+1}{7} = \frac{4}{7},\ F_1 = \frac{4}{7} = 0.57$$

$$\text{Macro: } P = \frac{\frac{3}{4}+0+\frac{1}{2}}{3} = \frac{5}{12},\ \ R = \frac{\frac{3}{3}+0+\frac{1}{2}}{3} = \frac{1}{2},\ F_1 = \frac{0.86+0+0.5}{3} = 0.45$$

$$\text{Weighted: } P = \frac{3}{7} \times \frac{3}{4} + \frac{2}{7} \times 0 + \frac{2}{7} \times \frac{1}{2} = \frac{13}{28},\ R = \frac{3}{7} \times \frac{3}{3} + \frac{2}{7} \times 0 + \frac{2}{7} \times \frac{1}{2} = \frac{4}{7},$$

$$F_1 = \frac{3}{7} \times 0.86 + \frac{2}{7} \times 0 + \frac{2}{7} \times 0.5 = 0.51$$

IRIS dataset

- Attribute Information:
  1. sepal length in cm
  2. sepal width in cm
  3. petal length in cm
  4. petal width in cm

Setosa :

Virginica :

Versicolour :

Iris dataset

5.1 , 3.8 , 1.6 , 0.2 ,	Iris-setosa
4.6 , 3.2 , 1.4 , 0.2 ,	Iris-setosa
5.3 , 3.7 , 1.5 , 0.2 ,	Iris-setosa
5.0 , 3.3 , 1.4 , 0.2 ,	Iris-setosa
7.0 , 3.2 , 4.7 , 1.4 ,	Iris-versicolor
6.4 , 3.2 , 4.5 , 1.5 ,	Iris-versicolor
6.9 , 3.1 , 4.9 , 1.5 ,	Iris-versicolor
5.5 , 2.3 , 4.0 , 1.3 ,	Iris-versicolor
6.5 , 2.8 , 4.6 , 1.5 ,	Iris-versicolor
5.7 , 2.8 , 4.5 , 1.3 ,	Iris-versicolor
7.2 , 3.0 , 5.8 , 1.6 ,	Iris-virginica
7.4 , 2.8 , 6.1 , 1.9 ,	Iris-virginica
7.9 , 3.8 , 6.4 , 2.0 ,	Iris-virginica
6.3 , 3.4 , 5.6 , 2.4 ,	Iris-virginica
6.4 , 3.1 , 5.5 , 1.8 ,	Iris-virginica
6.0 , 3.0 , 4.8 , 1.8 ,	Iris-virginica
6.9 , 3.1 , 5.4 , 2.1 ,	Iris-virginica

Reading

- E. Alpaydin, **Introduction to Machine Learning**, 4th ed., The MIT Press, 2020. (ch. 20)
- I. H. Witten, E. Frank. M. A. Hall, C. J. Pal, **Data Mining: Practical Machine Learning Tools and Techniques**. 4th ed., Morgan Kaufmann, 2017 (ch. 5)