Tutorial 2

Decision Tree, Cross-validation, Precision and Recall

Luke Chang

The University of Auckland

Mar. 2021

Objectives

1. Evaluation Metrics: Accuracy, Precision, Recall and F1 score
2. ROC curve and AUC
3. Should you trust the results?
4. Parametric Tests VS. Non-parametric Tests
5. Regression and Least Square Problem
6. Ensemble Methods

Confusion Matrix

Confusion Matrix can be applied to **binary** classification as well as for **multiclass** classification problems.

		Predicted: Positive	Predicted: Negative
Actual	**Positive**	True Positive	False Negative
	Negative	False Positive	True Negative

- True Positive (TP): Correctly classified.
- True Negative (TN): Correctly rejected.
- False Positive (FP): Incorrectly classified. Type I Error.
- False Negative (FN): Incorrectly rejected. Type II Error.

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

Confusion Matrix

How many selected items are relevant? Selected Elements $= \text{TP} + \text{FP}$

$$\text{Precision (P)} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

How many relevant items are selected? Relevant Elements $= \text{TP} + \text{FN}$

$$\text{Recall (R)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F_1 score is the **harmonic mean** between Precision and Recall.

$$F_1 = 2 \times \frac{P \times R}{P + R}$$

Example – Weather Prediction

Build a logistic regression model to predict the weather based on the humidity.
Recorded 10 days in total.

Class	Prediction
P	P
N	P
P	N
P	P
N	P
P	P
N	P
N	N
N	N
P	P

		Predicted P	Predicted N	Total
Actual	P	4	1	5
	N	3	2	5
	Total	7	3	**10**

$$\text{Acc.} = \frac{6}{10} = 0.6$$

$$\text{Precision (P)} = \frac{\text{TP}}{\text{TP}+\text{FP}} = \frac{4}{4+3} \approx 0.571$$

$$\text{Recall (R)} = \frac{\text{TP}}{\text{TP}+\text{FN}} = \frac{4}{4+1} \approx 0.8$$

$$F_1 = 2\frac{P \times R}{P+R} = 2 \times \frac{0.571 \times 0.8}{0.571+0.8} \approx 0.667$$

Caveat: A model with high Recall may also has high FPR (Type I Error).

Precision-Recall (PR) Curve (Optional)

Average precision (AP) summarizes such a plot as the weighted mean of precisions achieved at each threshold.

$$AP = \sum_n (R_n - R_{n-1}) P_n$$

- Where P_n and R_n are the precision and recall at the n-th threshold.
- A pair (P_n, P_k) is referred to as an *operating point*.

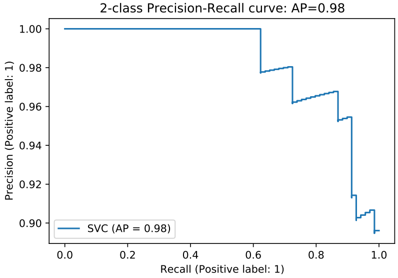


Figure: A SVM classifier trained on the Breast Cancer dataset

Receiver Operating Characteristic (ROC) Curve

- The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.
- Area Under Curve (AUC): The integration of the ROC function between 0 and 1.

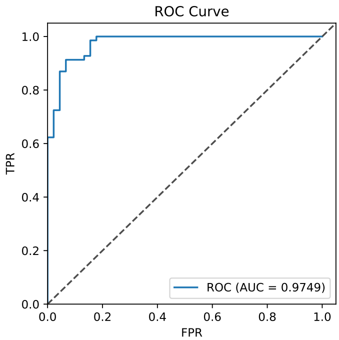


Figure: A SVM classifier trained on the Breast Cancer dataset

Example – Weather Prediction

Build a logistic regression model to predict the weather based on the humidity.
Recorded 10 days in total.

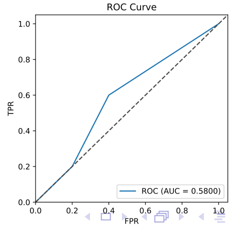
Class	Prediction	Thresholds 0	0.2	0.4	0.6	0.8	1
P	0.95	1	1	1	1	1	0
N	0.85	1	1	1	1	1	0
P	0.78	1	1	1	1	0	0
P	0.66	1	1	1	1	0	0
N	0.6	1	1	1	1	0	0
P	0.55	1	1	1	0	0	0
N	0.53	1	1	1	0	0	0
N	0.52	1	1	1	0	0	0
N	0.51	1	1	1	0	0	0
P	0.4	1	1	1	0	0	0

Counting TP and FP:

Threshold	0	0.2	0.4	0.6	0.8	1
TPR	1	1	1	0.60	0.2	0
FPR	1	1	1	0.4	0.2	0

Sort the results:

Threshold	0	0.2	0.4	0.6	0.8	1
TPR	0	0.2	0.6	1	1	1
FPR	0	0.2	0.4	1	1	1



Should you trust the results?

Scenario 1 from Page 48 in Week 2 slides

- I built a model based on the data you gave me
- It classified your data with 98% accuracy
- It should get 98% accuracy on the rest of your data

Should you trust them?

- They are reporting training error
- This might have nothing to do with test error
- E.g., They could have
  t a very deep decision tree

Why?

- If they only tried a few very simple models, the 98% might be reliable
- E.g., They only considered decision stumps with simple 1-variable rules

Should you trust the results?

Scenario 2 from Page 49 in Week 2 slides

- I built a model based on half of the data you gave me
- It classified the other half of the data with 98% accuracy
- It should get 98% accuracy on the rest of your data

Probably

- They computed the validation error once
- This is an unbiased approximation of the test error
- Trust them if you believe they did not violate the golden rule

Should you trust the results?

Scenario 4 from Page 51 in Week 2 slides

- I built 1 billion models based on half of the data you gave me
- One of them classified the other half of the data with 98% accuracy
- It should get 98% accuracy on the rest of your data

Probably not

- They computed the validation error a huge number of times
- They tried so many models, one of them is likely to work by chance

Why?

- If the 1 billion models were all extremely-simple, 98% might be reliable.

Regression and Least Square Problem

Ensemble Methods