Tutorial 3

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. Regression and Least Square Problem
5. Cross-Validation Questions
6. Ensemble Questions

Confusion Matrix

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

		Predicted	
		Positive	**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

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

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

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

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

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

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

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*.

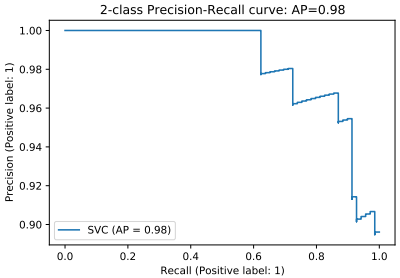


Figure: A SVM classifier trained on the Breast Cancer dataset

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.

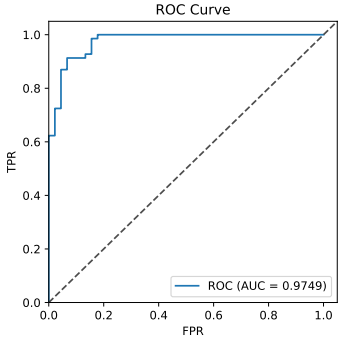


Figure: A SVM classifier trained on the Breast Cancer dataset

Example – Weather Prediction

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

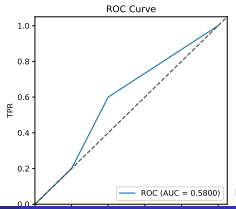
Class	Prediction	Thresholds 0	0.2	0.4	0.6	0.8	1
P	0.95	1	1	1	1	1	0
N	0.85	1	1	1	1	1	0
P	0.78	1	1	1	1	0	0
P	0.66	1	1	1	1	0	0
N	0.6	1	1	1	1	0	0
P	0.55	1	1	1	0	0	0
N	0.53	1	1	1	0	0	0
N	0.52	1	1	1	0	0	0
N	0.51	1	1	1	0	0	0
P	0.4	1	1	1	0	0	0

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	1	0.8	0.6	0.4	0.2	0
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

There are n samples, each sample has d features

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}, y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, x_i = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

X is a matrix which represents all samples. Such that:

$$\hat{y} = Xw = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nd} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix}$$

Mean Squared Error:

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

Cross Validation Questions

Question 1

If you do 2-fold cross validation, 10-fold cross validation or leave-one-out, or use a 70/30 percent train/validation single split. What effect will this have on your results?

Question 2

Which will give you the best representative value to the "unseen test set"?

Cross Validation Questions

Question 1

If you do 2-fold cross validation, 10-fold cross validation or leave-one-out, or use a 70/30 percent train/validation single split. What effect will this have on your results?

- Leave-one-out means that you have a bigger training set and a bigger validation set. Also you have N repetitions where N is the size of your dataset.
- 2-fold gives you a much smaller training set but a bigger validation set. A bigger validation set is good but a smaller training set is not – high bias.
- 10-fold validation gives you a bigger training set but an even smaller validation set than 2-fold – high variance.

Question 2

Which will give you the best representative value to the "unseen test set"?

Cross Validation Questions

Question 1

If you do 2-fold cross validation, 10-fold cross validation or leave-one-out, or use a 70/30 percent train/validation single split. What effect will this have on your results?

- Leave-one-out means that you have a bigger training set and a bigger validation set. Also you have N repetitions where N is the size of your dataset.
- 2-fold gives you a much smaller training set but a bigger validation set. A bigger validation set is good but a smaller training set is not – high bias.
- 10-fold validation gives you a bigger training set but an even smaller validation set than 2-fold – high variance.

Question 2

Which will give you the best representative value to the "unseen test set"?

Probably leave-one-out, with 2-fold the training set might be too small and with 10-fold then validation set might be too small, but tenfold is typically better than 2-fold.

Cross Validation Questions

Question 3

Does it matter how large your original dataset is?
Will you get a different answer for a very big or a very small dataset?

Question 4

Which will be the fastest in terms of computation?

Cross Validation Questions

Question 3

Does it matter how large your original dataset is?
Will you get a different answer for a very big or a very small dataset?

- For a very big dataset – all will probably give you the same answer
- For a very small dataset – you would probably have to do 10-fold or leave-one-out to get good results

Question 4

Which will be the fastest in terms of computation?

Cross Validation Questions

Question 3

Does it matter how large your original dataset is?
Will you get a different answer for a very big or a very small dataset?

- For a very big dataset – all will probably give you the same answer
- For a very small dataset – you would probably have to do 10-fold or leave-one-out to get good results

Question 4

Which will be the fastest in terms of computation?

- Leave-one-out is the most expensive, you have to learn N models.
- 10-fold CV means you need to learn 10 models and 2-fold means you have to learn 2 models.

Ensemble Questions

Question 1

What are the two key factors an ensemble must have?

Question 2

Bagging changes two things in the dataset, what are they?

Ensemble Questions

Question 1

What are the two key factors an ensemble must have?

- Each tree must perform better than random guess/ average.
- Must be uncorrelated.

Question 2

Bagging changes two things in the dataset, what are they?

Ensemble Questions

Question 1

What are the two key factors an ensemble must have?

- Each tree must perform better than random guess/ average.
- Must be uncorrelated.

Question 2

Bagging changes two things in the dataset, what are they?

- The choice of data instances
- The distribution over them (With replacement)

Ensemble Questions

To apply binary classification method on multi-class data, you need:

- One-Vs-Rest (OVR)
- One-Vs-One (OVO)
- Error-Correcting Output Code (ECOC)

Question 3

Does Error-Correcting Output Code (ECOC) work better when there are many classes or when there are few? Why?

Question 4

If you are going to choose with replacement can you then make your training set as big as you want? Will this work?

Ensemble Questions

To apply binary classification method on multi-class data, you need:

- One-Vs-Rest (OVR)
- One-Vs-One (OVO)
- Error-Correcting Output Code (ECOC)

Question 3

Does Error-Correcting Output Code (ECOC) work better when there are many classes or when there are few? Why?

It will work better with many classes with 3 classes there are very few codes.
Basic concept: Encode the N-output classes with number of bits.

Question 4

If you are going to choose with replacement can you then make your training set as big as you want? Will this work?

Ensemble Questions

To apply binary classification method on multi-class data, you need:

- One-Vs-Rest (OVR)
- One-Vs-One (OVO)
- Error-Correcting Output Code (ECOC)

Question 3

Does Error-Correcting Output Code (ECOC) work better when there are many classes or when there are few? Why?

It will work better with many classes with 3 classes there are very few codes.
Basic concept: Encode the N-output classes with number of bits.

Question 4

If you are going to choose with replacement can you then make your training set as big as you want? Will this work?

You cannot make a million instances out of 100, but if you are short on data you can use this to try and bolster your results – especially in ensembles.

Ensemble Questions

Question 5

What is one of the main differences between random forest and bagging?

- What will the effect be of having a dataset with a larger or smaller number of instances?
- What will the effect be of having a dataset with a larger or smaller number of features?

Ensemble Questions

Question 5

What is one of the main differences between random forest and bagging? Random forest samples the features.

- What will the effect be of having a dataset with a larger or smaller number of instances?
  The effect on bagging and random forest will be the same – they both sample the instance space with replacement.
- What will the effect be of having a dataset with a larger or smaller number of features?
  Since random forests sample the features you might get better results when there are a lot of features because you got rid of a lot of noise – but with a data set with only a few features you might do worse because you are not left with enough features to make a good classifier

Ensemble Questions

Question 6

Will variable importance in Random Forest always give you the "correct" answer? Why or why not?

Ensemble Questions

Question 6

Will variable importance in Random Forest always give you the "correct" answer? Why or why not?

No because if you have correlated attributes Random forest will say neither are important even if they are the most important.
This is because it randomizes the variables one at a time, thereby relying on the correlated variable when each is randomized.
Example Logic Function: $A \vee (B \wedge C)$ - B and C are correlated. If we train a tree with only A and B, or A and C, we will not have the correct logic.

Ensemble Questions

Question 7

Which of the "methods for constructing ensembles" does random forest use?

Question 8

Which of the "methods for constructing ensembles" does XGBoost use?

Question 9

What is one of the main differences between XG Boost and Random Forests?

Ensemble Questions

Question 7

Which of the "methods for constructing ensembles" does random forest use?

Manipulating the training set, Manipulating the input features (columns), Injecting Randomness

Question 8

Which of the "methods for constructing ensembles" does XGBoost use?

Question 9

What is one of the main differences between XG Boost and Random Forests?

Ensemble Questions

Question 7

Which of the "methods for constructing ensembles" does random forest use?

Manipulating the training set, Manipulating the input features (columns), Injecting Randomness

Question 8

Which of the "methods for constructing ensembles" does XGBoost use?

Manipulating the training set, Manipulating the input features (columns), Injecting Randomness

Question 9

What is one of the main differences between XG Boost and Random Forests?

Ensemble Questions

Question 7

Which of the "methods for constructing ensembles" does random forest use?

Manipulating the training set, Manipulating the input features (columns), Injecting Randomness

Question 8

Which of the "methods for constructing ensembles" does XGBoost use?

Manipulating the training set, Manipulating the input features (columns), Injecting Randomness

Question 9

What is one of the main differences between XG Boost and Random Forests?

XG Boost chooses without replacement so does not change the distribution of the dataset – also Boosting will have to use a smaller training set by definition; also RF uses democratic voting and XG Boost uses weighted voting

Ensemble Questions

Question 10

What is the difference between attribute error and class error and does it have a major effect in ensembles?

Question 10a

Are some ensemble methods more sensitive to this?

Ensemble Questions

Question 10

What is the difference between attribute error and class error and does it have a major effect in ensembles?

Attribute error is when noises is added to an input variable X and class error is when noise is added to the class Y

Question 10a

Are some ensemble methods more sensitive to this?

All ensemble methods should work will with attribute error to a point.
But AdaBoost is very sensitive to class errors because it weights instances, whereas XGBoost weights trees.

Ensemble Questions

Boosting algorithm increases the weights of miss-classified data points over time, so that the next classifier will pay extra attention to get them right.

Question 11

What is the main difference between AdaBoost and XGBoost?

Ensemble Questions

Boosting algorithm increases the weights of miss-classified data points over time, so that the next classifier will pay extra attention to get them right.

Question 11

What is the main difference between AdaBoost and XGBoost?

- AdaBoost weights the data instances, and XGBoost weights the trees added into the ensemble – non-democratic voting.
- AdaBoost uses "Weak learner". Each tree has only 1 node and 2 leaves (a stump).
- XGBoost uses larger tree similar to Random Forest, and uses `max_depth` as a hyperparameter.