INFO-2301

Michael Paul

April 10, 2017

Prediction

Classification

- Assign a discrete value y to input **x**
- The possible values of y are called **classes**

Regression

- Assign a continuous value y to input **x**

Reminder: **x** is usually a vector

- The dimensions of **x** correspond to **features**
- Features are properties like word counts, pixel values, etc.

Evaluation

Suppose you build a classifier or regression model.

How do you measure how well it performs?

Today: we'll look at a variety of metrics for measuring performance, and discuss how to use them

sklearn

Scoring	Function
Classification	
'accuracy'	`metrics.accuracy_score`
'average_precision'	`metrics.average_precision_score`
'f1'	`metrics.f1_score`
'f1_micro'	`metrics.f1_score`
'f1_macro'	`metrics.f1_score`
'f1_weighted'	`metrics.f1_score`
'f1_samples'	`metrics.f1_score`
'neg_log_loss'	`metrics.log_loss`
'precision' etc.	`metrics.precision_score`
'recall' etc.	`metrics.recall_score`
'roc_auc'	`metrics.roc_auc_score`
Clustering	
'adjusted_rand_score'	`metrics.adjusted_rand_score`
Regression	
'neg_mean_absolute_error'	`metrics.mean_absolute_error`
'neg_mean_squared_error'	`metrics.mean_squared_error`
'neg_median_absolute_error'	`metrics.median_absolute_error`
'r2'	`metrics.r2_score`

http://scikit-learn.org/stable/modules/model_evaluation.html

Evaluation: Classification

How often does the predicted class match the true class?

Evaluation: Classification

Accuracy: % of predictions that are correct

- A simple metric that is easy to understand in many cases
- Misleading if the distribution of classes is imbalanced
  - If 95% of data instances are the same class, then you can always predict that one class and get 95% accuracy. Not very informative.

Evaluation: Classification

$$\text{Precision:} \quad \frac{\text{\# classified positive that are actually positive}}{\text{\# classified positive}}$$

- Related to accuracy, but it's the accuracy among only the instances your classifier predicted to be 'positive'

Evaluation: Classification

$$\text{Recall:} \quad \frac{\text{\# positive instances that were classified positive}}{\text{\# positive instances}}$$

- How much does your classifier capture?
- Usually in conflict with precision
  - You can increase recall by classifying more instances as positive, but that might drop your precision

Evaluation: Classification

Precision vs recall: which is more important? Depends on task.

- Spam classification: important not to misclassify legitimate email as spam, so high precision is necessary (even if recall drops)
- Search engines: important to grab as much as possible that matches a query (high recall), mistakes aren’t a big deal because user can ignore

Evaluation: Classification

F-score: a type of average between precision and recall

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

This is called the harmonic mean – both need to be high for F to be high

Evaluation: Classification

Note that precision/recall/F-score assume there is a ‘positive’ and ‘negative’ class (need to define)

What if you have more than two labels?

- Accuracy is always an option
- Can calculate precision/recall/F-score for each class, and average the scores

Evaluation: Regression

No longer makes sense to ask, “was the prediction correct?”

Instead: how *close* was the prediction?

Evaluation: Regression

Mean squared error: this is what least squares regression will minimize

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

Mean absolute error: you saw this in your regression assignment

$$\frac{1}{N}\sum_{i=1}^{N}|x_i - y_i|$$

Evaluation: Regression

r^2: the square of the correlation between the predicted values and the true values

- 0 means no correlation, 1 means perfect correlation

Equivalent definition:

- The total sum of squares (proportional to the variance of the data):

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2,$$

- The regression sum of squares, also called the explained sum of squares:

$$SS_{\text{reg}} = \sum_i (f_i - \bar{y})^2,$$

- The sum of squares of residuals, also called the residual sum of squares:

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

The most general definition of the coefficient of determination is

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}.$$

Evaluation: Regression

Which metric to use?

Mean error

- Easy to interpret
- Depends on units

r^2

- Does not depend on units or scale, so can be compared across datasets/tasks

Validation Data

Easy to measure how well your model does on the data you gave it, but how can you estimate how well it will perform on new data in the future?

- Key: don’t evaluate it on the same data you use to build the model

Validation Data

Training data (in-sample) vs test data (out-of-sample)

- Usually remove 10-20% of your dataset as a “held-out” set for testing

Cross-validation: split your dataset into several “folds”, train on all but one, test on the remaining one, then repeat and average the results

- 5-fold cross-validation:
  - Split your data into 5 smaller datasets
  - Train on 4/5, test on the remaining 1/5
  - Repeat this 5 different times, so each fold is the test set once
  - Average the 5 results (precision/recall/F)

Overfitting

If your performance is *too* good on the training data, you can actually hurt the performance on future unseen data

- Your model might pick up idiosyncrasies in the training data that do not generalize to other data
- This is why it's important to evaluate on held-out test data