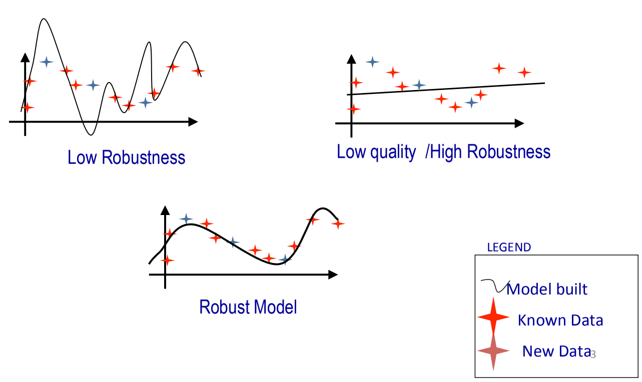
Model Selection & Evaluation

Many of these slides are derived from Seyong Kim, Ziv-Bar Joseph. Thanks!

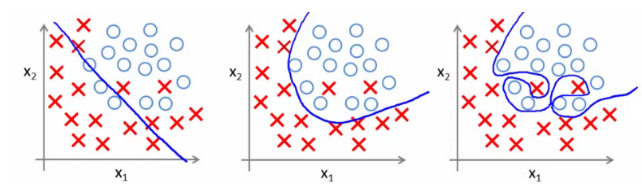
Recall: overfitting

- What is overfitting?
- Why does it happen?
- How can we avoid it?

What is a good model?



Over vs Under fitting



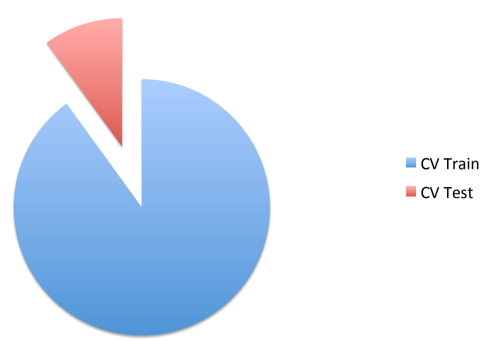
UNDERFITTING (high bias)

OVERFITTING (high variance)

Model Selection

- Suppose we are trying to select among several different models for a learning problem.
- E.g.
  - Full tree vs. tree pruned to depth 5 vs. random forest?

Cross Validation



Practical issues for CV

- How to big of a slice of the pie?
  - Commonly used K = 10 folds (thus each fold is 10% of the data)
  - LOOCV (K=N, number of training instances)

- One important point is that the test data is never used in CV (only the training data), because doing so would result in overly (indeed dishonest) optimistic accuracy rates during the testing phase.

- Stratification – should you balance the classes across the folds?

7

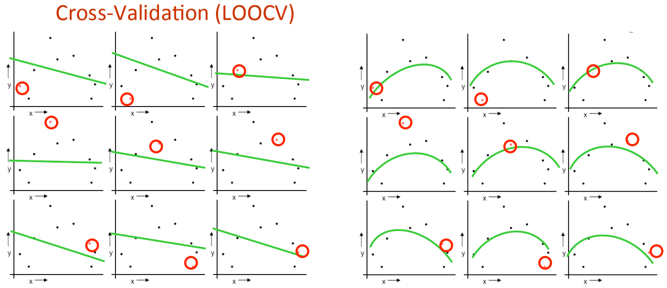
How to handle your data

- Two regimes:
  1. Cross Validation
     - Split into Train/Test (e.g. 70,30%)
     - Perform cross validation on training data to set parameters or choose an algorithm
     - Report final accuracy by training on all of training data (with your final chosen parameters) and predicting on test data
  2. Validation Set
     - Split into Train/Validation/Test (e.g. 70,10,20%)
     - Train on Training data, test on validation to set parameters or choose an algorithm
     - Report final accuracy by training on all of training data (with your final chosen parameters) and predicting on test data.

8

Example:

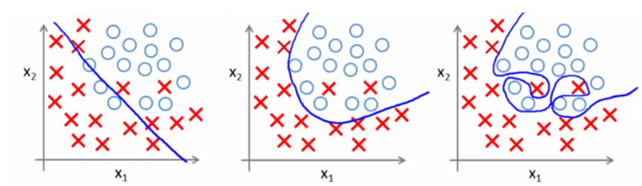
- When $\alpha=1/N$, the algorithm is known as Leave-One-Out-Cross-Validation (LOOCV)



$MSE_{LOOCV}(M_1)=2.12$ $MSE_{LOOCV}(M_2)=0.962$

9

Why is CV so important?



UNDERFITTING (high bias)

OVERFITTING (high variance)

10

Other Evaluation Methods

- Random subsampling / Monte Carlo cross validation
  - choose a test set randomly and repeatedly
  - like cross-validation except test sets need not be disjoint
- Bootstrap
  - choose a test set randomly with replacement
  - like random sampling, but with replacement
  - Pessimistic estimate, corrected with .632 bootstrap estimate

More info: http://web.cs.iastate.edu/~jtian/cs573/Papers/Kohavi-IJCAI-95.pdf

11

Measuring Performance

- We usually calculate performance on test data
- Calculating performance on training data is called resubstitution and is an *optimistic* measure of performance
  - why?
  - because it can't detect overfitting

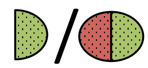
12

Measuring Performance (Classification)

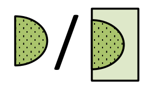
- Accuracy
  - (# test instances correctly labeled)/(# test instances)
- Error
  - 1- accuracy
  - (# test instances incorrectly labeled)/(# test instances)

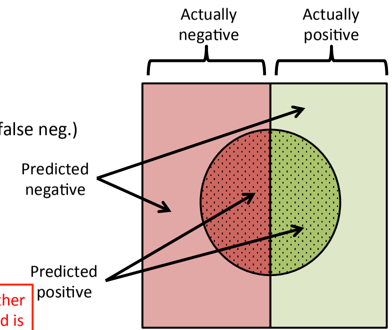
Measuring Performance (Classification)

- Precision
  - true positives/(true pos. + false pos.)



- Recall
  - true positives/(true pos. + false neg.)





How to remember: the first word is whether the prediction is correct, the second word is what the prediction was.

Measuring Performance (Classification)

- F1
  - harmonic mean of precision and recall
  - 2*(p*r)/(p+r)

Measuring Performance (Regression)

- Regression
  - predicting a real number
- Root Mean Squared Error (RMSE)
  - sometimes just MSE (no sqrt)

$$\sqrt{\frac{1}{N}\underbrace{\sum_{i=1}^{N}}_{\text{sum over all N test instances}}\underbrace{(\hat{y}_i - y_i)}_{\text{error}}^2}$$

Performance

- Comparing performance of classifiers
- How do you know if your accuracy number is “high” or error is “low”?

Exercise

Consider the task of building a classifier from random data, where the attribute values are generated randomly irrespective of the class labels. Assume the data set contains records from two classes, “+” and “–.” Half of the data set is used for training while the remaining half is used for testing.

(a) Suppose there are an **equal number** of positive and negative records in the data and the **classifier predicts every test record to be positive**. What is the expected error rate of the classifier on the test data?

Answer: 50%.

(b) Repeat the previous analysis assuming that the classifier predicts each test record to be **positive class** with probability 0.8 and **negative class** with probability 0.2.

Answer: 50%.

Exercise

Consider the task of building a classifier from random data, where the attribute values are generated randomly irrespective of the class labels. Assume the data set contains records from two classes, "+" and "−." Half of the data set is used for training while the remaining half is used for testing.

(c) Suppose **2/3** of the data belong to the **positive** class and the remaining **1/3** belong to the **negative** class. What is the **expected error** of a classifier that **predicts every test record to be positive**?

Answer: (2/3)*0+(1/3)*1 = 33%.

(d) Repeat the previous analysis assuming that the classifier predicts each test record to be positive class with probability 2/3 and negative class with probability 1/3.

Answer: (2/3)*(1/3)+(1/3)*(2/3) = 44.4%.

Exercise

Consider a classifier X that has **Accuracy = 50%** on a (test) dataset with a class taking 2 possible values (A, B).

The distribution of the instances for each class value is:

A:50,   B:50.

How does X compare to a random classifier Y that outputs A, and B, 50%, 50% of the time, respectively.

Answer:

Y's accuracy:

(50*50/100 + 50*50/100)/100 = 50%

- So, X performs the same (accuracy-wise) as Y.

Exercise

Consider a classifier X that has **Accuracy = 50%** on a (test) dataset with a class taking 4 possible values (A, B, C, and D).

The distribution of the instances for each class value is

A:25,   B:25,   C:25,   and   D:25.

How does X compare to a random classifier Y that outputs A, B, C, and D 25%, 25%, 25%, and 25% of the time, respectively.

Answer:

Y's accuracy:

(25*25/100 + 25*25/100 + 25*25/100 + 25*25/100)/100 = 25%

- So, X does twice better than Y (accuracy-wise).

Exercise

The distribution of the instances for each class value is A:25, B: 25, C:25, and D:25.

Random classifier Y outputs A, B, C, and D, 25%, 25%, 25%, and 25% of the time, respectively.

Precision and Recall (wrt A)?

Answer:

Y will say 25% of the time "A" and 75% of the time "not A".

TP = 1/4*1/4, FP = 3/4*1/4, FN = 1/4*3/4

Precision=

TP/(TP+FP) = 25%

Recall=

TP/(TP+FN) = 25%

Exercise

The distribution of the instances for each class value is A:10, B: 40, C:25, and D:25.

Random classifier Y outputs A, B, C, and D, 50%, 30%, 10%, and 10% of the time, respectively.

Precision and Recall (wrt A)?

Answer:

Y will say 50% of the time "A" and 50% of the time "not A".

TP = ? FP = ? FN = ?

Precision=

TP/(TP+FP) = 1/10= 10%

Recall=

TP/(TP+FN) = 1/2= 50%