Experimentation and Evaluation

Instructor: Jessica Wu -- Harvey Mudd College

The instructor gratefully acknowledges Eric Eaton (UPenn), David Kauchak (Pomona), and the many others who made their course materials freely available online.

Evaluation Basics

Learning Goals

- Describe problems with using training set to evaluate performance
- Define metrics for evaluating performance
  - accuracy, error
  - confusion matrix
  - sensitivity, specificity
  - receiver operating curve (ROC)

Comparing Classifiers

Given: two classifiers, C_1 and C_2

Goal: choose the best one to use for future predictions

C_1 and C_2 may be

- same learning model with **different complexities** or **hyperparameters**
  - decision trees : different depths
  - *k*-NN : different choices of *k*
- different learning models

Can we use training accuracy to choose between them?

- No!
  e.g., C_1 = pruned decision tree, C_2 = 1-NN
  training_accuracy(1-NN) = 100% but may not be best
- Instead, choose based on test accuracy...

Based on slides by Piyush Rai and Eric Eaton (originally by Padhraic Smyth)

Training Data and Test Data

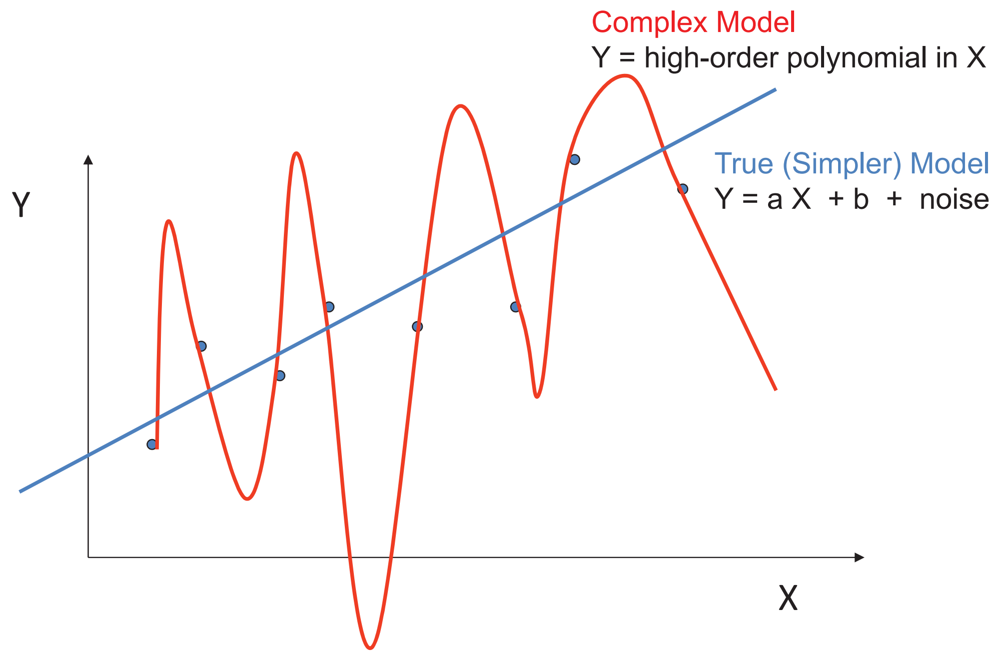
- Training data: data used to build model
- Test data: new data, not used in training process

- Training performance is often poor indicator of generalization performance
  - generalization is what we <u>really</u> care about in ML
  - easy to overfit to training data
  - performance on test data is good indicator of generalization performance

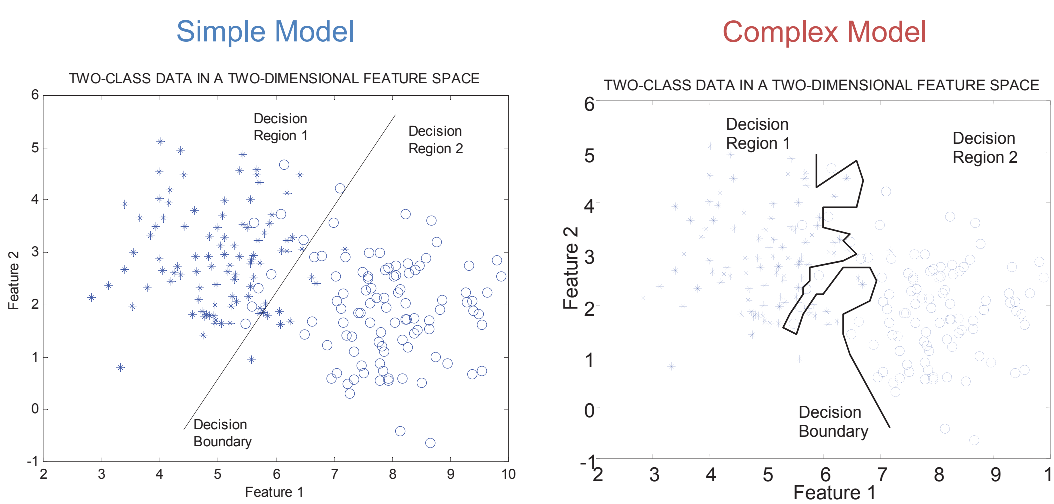
- Test accuracy more important than training accuracy

Based on slide by Eric Eaton

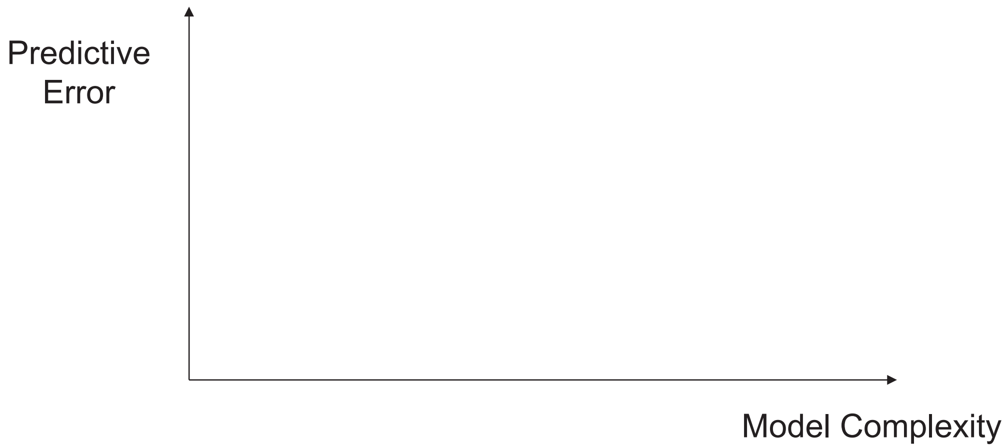
Example: The Overfitting Phenomenon



Example: The Overfitting Phenomenon



How Overfitting Affects Prediction

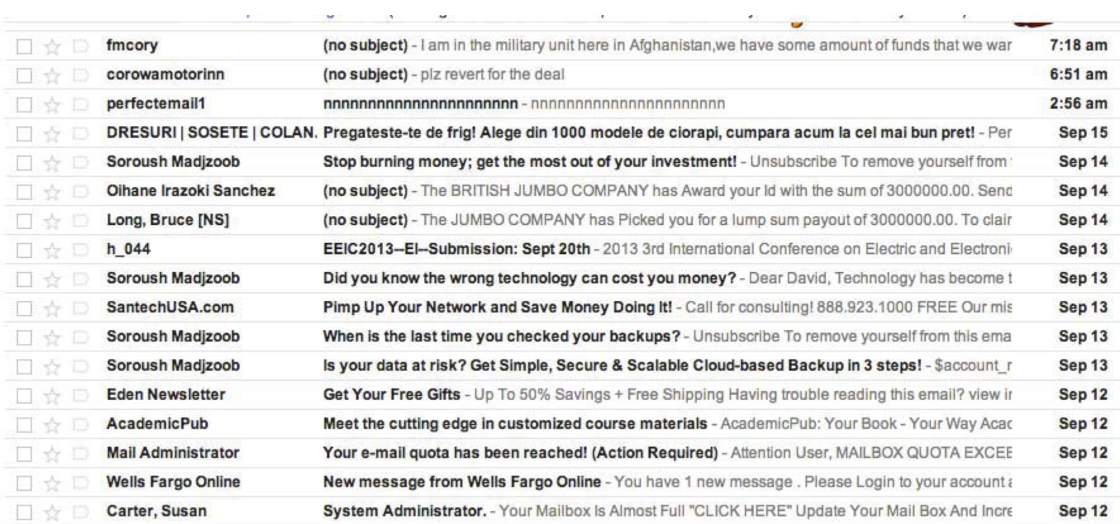


Based on slide by Padhraic Smyth, UCIrvine

Real-world classification

But how do we get test data?

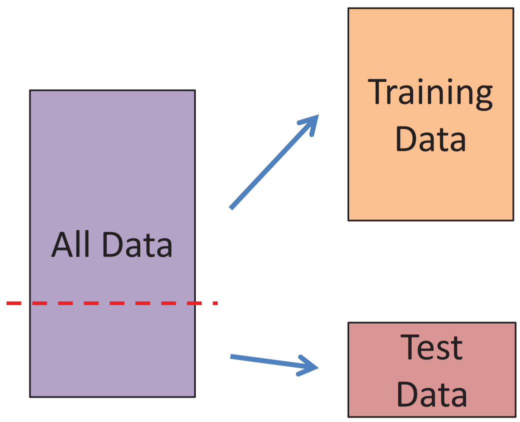
Google has labeled training data, for example from people clicking “spam” button, but when new messages come in, they are not labeled



fmcory	**(no subject)** - I am in the military unit here in Afghanistan,we have some amount of funds that we war	**7:18 am**
corowamotorinn	**(no subject)** - plz revert for the deal	**6:51 am**
perfectemail1	**nnnnnnnnnnnnnnnnnnnnnn** - nnnnnnnnnnnnnnnnnnnnnn	**2:56 am**
DRESURI | SOSETE | COLAN.	**Pregateste-te de frig! Alege din 1000 modele de ciorapi, cumpara acum la cel mai bun pret!** - Per	**Sep 15**
Soroush Madjzoob	**Stop burning money; get the most out of your investment!** - Unsubscribe To remove yourself from	**Sep 14**
Oihane Irazoki Sanchez	**(no subject)** - The BRITISH JUMBO COMPANY has Award your Id with the sum of 3000000.00. Send	**Sep 14**
Long, Bruce [NS]	**(no subject)** - The JUMBO COMPANY has Picked you for a lump sum payout of 3000000.00. To clair	**Sep 14**
h_044	**EEIC2013--EI--Submission: Sept 20th** - 2013 3rd International Conference on Electric and Electroni	**Sep 13**
Soroush Madjzoob	**Did you know the wrong technology can cost you money?** - Dear David, Technology has become t	**Sep 13**
SantechUSA.com	**Pimp Up Your Network and Save Money Doing It!** - Call for consulting! 888.923.1000 FREE Our mis	**Sep 13**
Soroush Madjzoob	**When is the last time you checked your backups?** - Unsubscribe To remove yourself from this ema	**Sep 13**
Soroush Madjzoob	**Is your data at risk? Get Simple, Secure & Scalable Cloud-based Backup in 3 steps!** - $account_r	**Sep 13**
Eden Newsletter	**Get Your Free Gifts** - Up To 50% Savings + Free Shipping Having trouble reading this email? view ir	**Sep 12**
AcademicPub	**Meet the cutting edge in customized course materials** - AcademicPub: Your Book - Your Way Acac	**Sep 12**
Mail Administrator	**Your e-mail quota has been reached! (Action Required)** - Attention User, MAILBOX QUOTA EXCEE	**Sep 12**
Wells Fargo Online	**New message from Wells Fargo Online** - You have 1 new message . Please Login to your account a	**Sep 12**
Carter, Susan	**System Administrator.** - Your Mailbox Is Almost Full "CLICK HERE" Update Your Mail Box And Incre	**Sep 12**

Based on slide by David Kauchak

Classification Evaluation

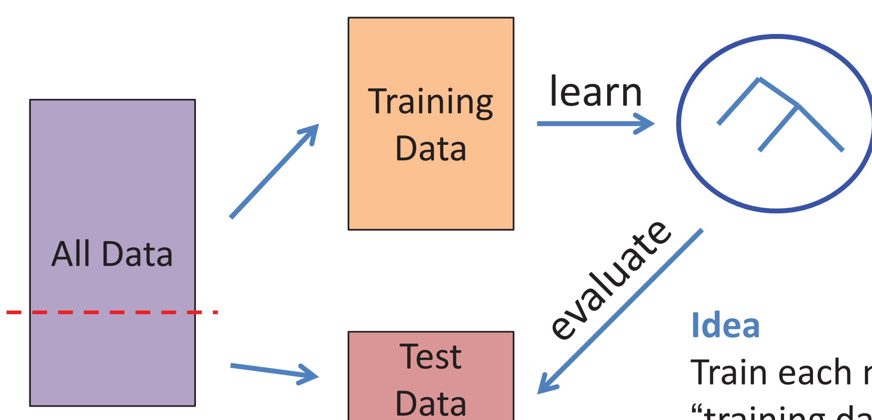


Use labeled data we have already to create test set with known labels!

Why can we do this?
Remember, we assume there's an underlying distribution that generates both training and test examples

Based on slides by Eric Eaton and David Kauchak

Classification Evaluation

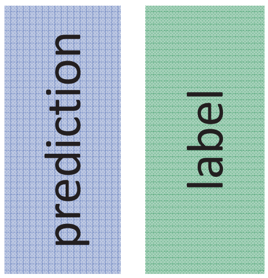


Idea
Train each model on "training data"...
...and then test each model on test data

Based on slide by Eric Eaton

Evaluation Metrics

To evaluate model, compare predicted labels to actual labels



Accuracy: proportion of examples where we predicted correct label

$$\text{accuracy} = \frac{\#\ \text{correct predictions}}{\#\ \text{test instances}}$$

Error: proportion of examples where we predicted incorrect label

$$\text{error} = 1 - \text{accuracy}$$

$$= \frac{\#\ \text{incorrect predictions}}{\#\ \text{test instances}}$$

Based on slide by David Kauchak

Confusion Matrices

How can we understand what types of mistakes a classifier makes?

activity recognition from video

Riemannian metric with Binet Cauchy kernel, Leave−one−out 1−NN

actual class \ predicted class	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	100	0	0	0	0	0	0	0	0	0
jack	0	100	0	0	0	0	0	0	0	0
jump	0	0	89	0	0	0	11	0	0	0
pjump	0	0	0	100	0	0	0	0	0	0
run	0	0	0	0	89	0	11	0	0	0
side	0	0	0	0	0	100	0	0	0	0
skip	0	0	0	0	0	0	100	0	0	0
walk	0	0	0	0	0	0	0	100	0	0
wave1	0	0	0	0	0	0	0	0	67	33
wave2	0	0	0	0	0	0	0	0	0	100

actual class

predicted class

Based on slide by David Page
[Image source: jhu.vision.edu]

Confusion Matrix for 2-class problems

- Imagine a classifier that identifies presence of disease

		predicted class	
		Yes	No
actual class	Yes	TP	FN
	No	FP	TN

$$\text{accuracy} = \frac{TP + TN}{P + N}$$

TP = person tests positive and really has disease
TN = person tests negative and really does not have disease
FP = person tests positive and does not have disease
FN = person tests negative and has disease

P = actual class is positive = TP + FN
N = actual class is negative = TN + FP

Based on slide by Eric Eaton

Is Accuracy an Adequate Measure?

Confusion Matrix

- Given dataset of P positive instances and N negative instances:

actual class \ predicted class	Yes	No
Yes	TP	FN
No	FP	TN

$$\text{accuracy} = \frac{TP + TN}{P + N}$$

- Imagine a classifier that identifies presence of disease

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

(true positive rate) = probability of positive test given person has disease

$$\text{specificity} = \frac{TN}{TN + FP}$$

(true negative rate) = probability of negative test given person does not have disease

Confusion Matrix: Cancer Dataset

patients with cancer \ screen test	Yes	No
Yes	20	10
No	180	1820

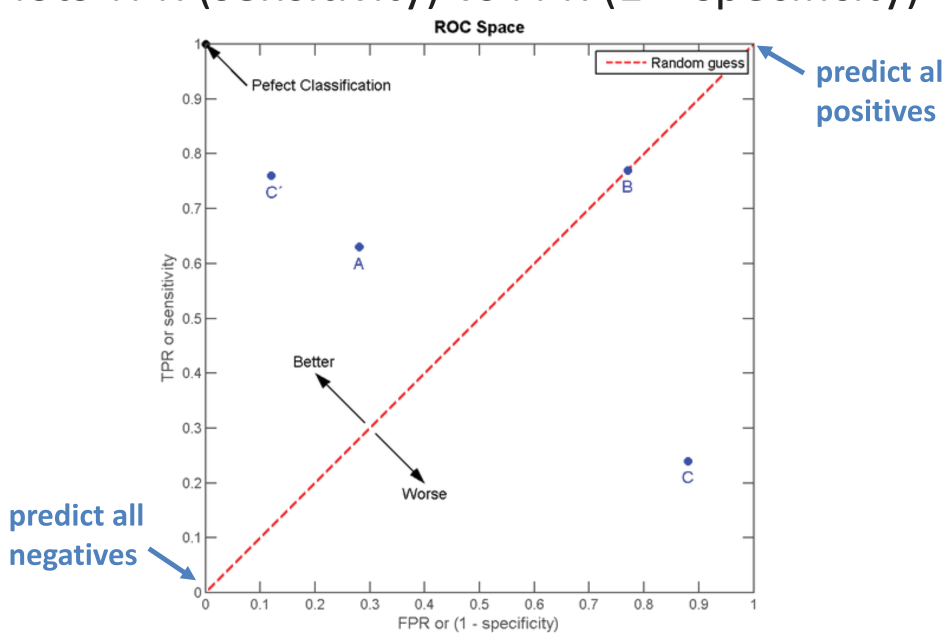
Compute accuracy, sensitivity, and specificity

Limitations of Sensitivity / Specificity

- How can you maximize sensitivity [= TP / (TP + FN)]?
- How can you maximize specificity [= TN / (TN + FP)]?
- What does this mean?

Receiver Operating Characteristic

- Plots TPR (sensitivity) vs FPR (1 – specificity)



[Source: Wikipedia]