Learning theory
Lecture 8

David Sontag
New York University

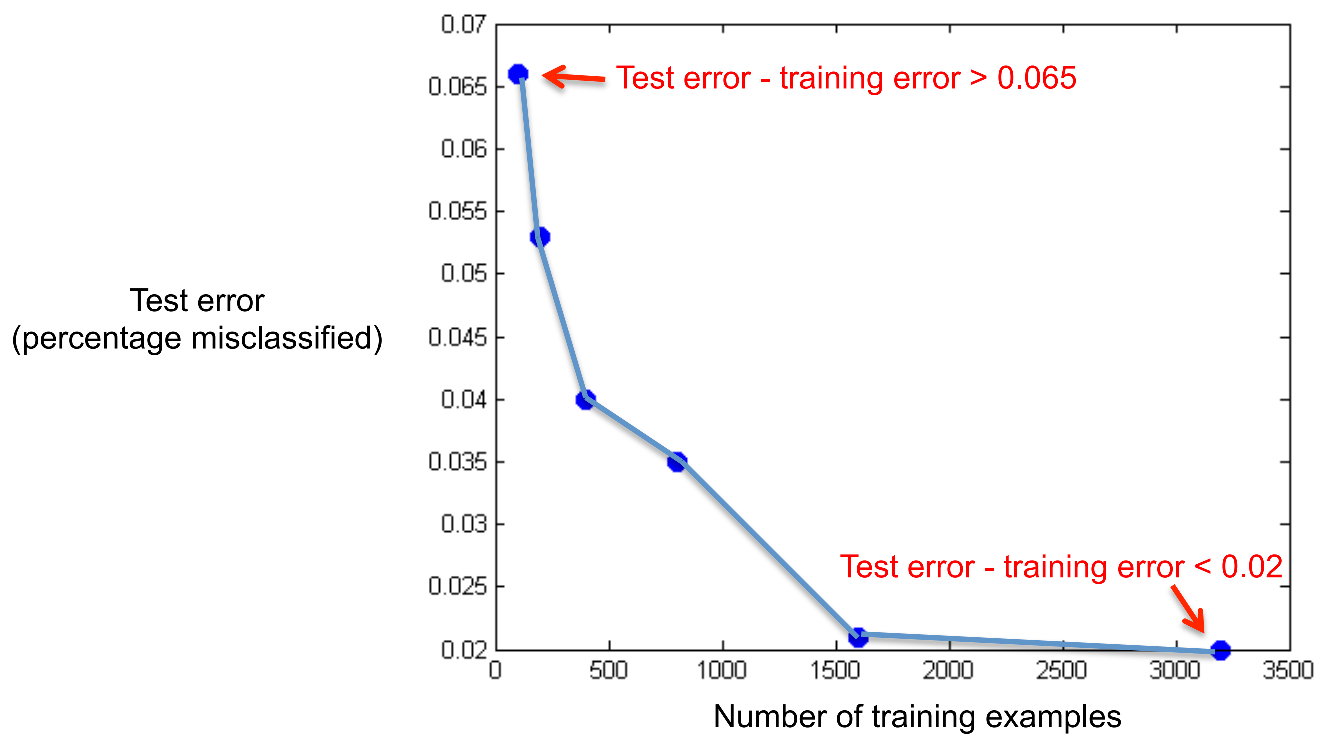
Slides adapted from Carlos Guestrin & Luke Zettlemoyer

What's next…

- We gave several machine learning algorithms:
  - Perceptron
  - Linear support vector machine (SVM)
  - SVM with kernels, e.g. polynomial or Gaussian
- How do we guarantee that the learned classifier will perform well on test data?
- How much training data do we need?

Example: Perceptron applied to spam classification

- In your homework 1, you trained a spam classifier using perceptron
  - The training error was always zero
  - With few data points, there was a big gap between training error and test error!



How much training data do you need?

- Depends on what *hypothesis class* the learning algorithm considers

- For example, consider a memorization-based learning algorithm
  - Input: training data $S = \{ (\mathbf{x}_i, y_i) \}$
  - Output: function $f(\mathbf{x})$ which, if there exists $(\mathbf{x}_i, y_i)$ in S such that $\mathbf{x}=\mathbf{x}_i$, predicts y_i, and otherwise predicts the majority label
  - This learning algorithm will always obtain zero training error
  - But, it will take a ***huge*** amount of training data to obtain small test error (i.e., its generalization performance is horrible)

- Linear classifiers are powerful precisely because of their simplicity
  - Generalization is easy to guarantee

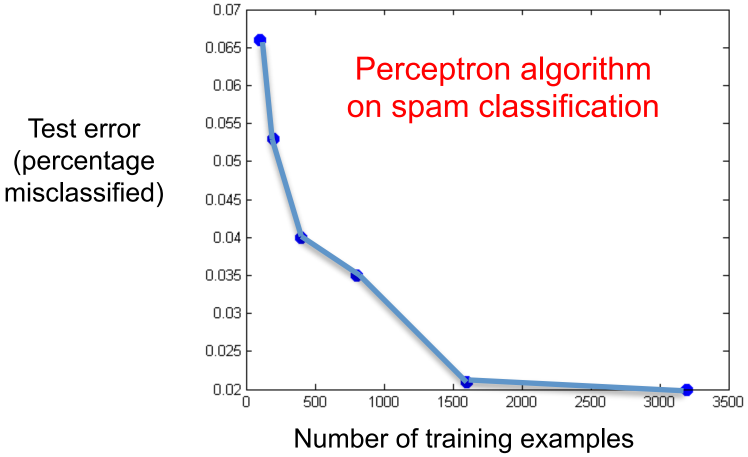
Roadmap of next two lectures

1. Generalization of finite hypothesis spaces

2. VC-dimension

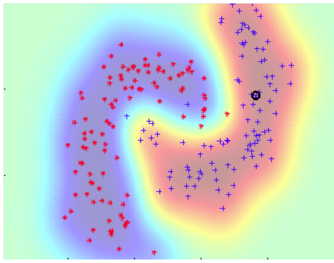
   - Will show that linear classifiers need to see approximately **d** training points, where **d** is the dimension of the feature vectors

   - Explains the good performance we obtained using perceptron!!!! (we had a few thousand features)



3. Margin based generalization

   - Applies to **infinite** dimensional feature vectors (e.g., Gaussian kernel)

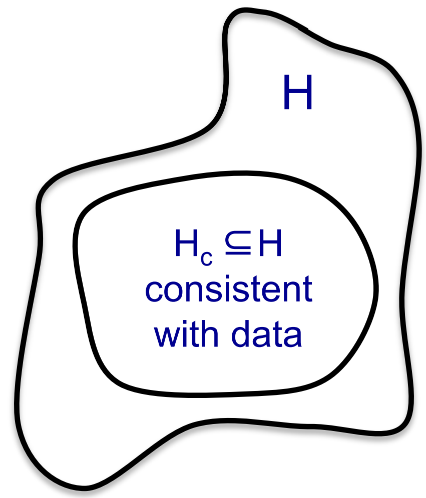


[Figure from Cynthia Rudin]

How big should your validation set be?

- In PS1, you tried many configurations of your algorithms (avg vs. regular perceptron, max # of iterations) and chose the one that had smallest validation error
- Suppose in total you tested **|H|=**40 different classifiers on the validation set of **m** held-out e-mails
- The best classifier obtains 98% accuracy on these **m** e-mails!!!
- But, what is the true classification accuracy?
- How large does **m** need to be so that we can guarantee that the best configuration (measured on validate) is truly good?

A simple setting…

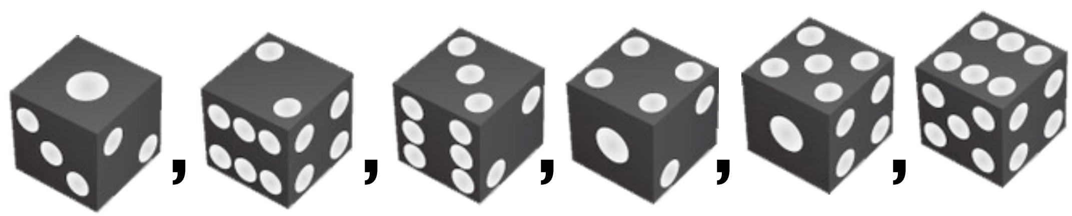


- Classification
  - m data points
  - **Finite** number of possible hypothesis (e.g., 40 spam classifiers)
- A learner finds a hypothesis *h* that is **consistent** with training data
  - Gets zero error in training: $error_{train}(h) = 0$
  - I.e., assume for now that one of the classifiers gets 100% accuracy on the **m** e-mails (we'll handle the 98% case afterward)
- What is the probability that *h* has more than ε **true** error?
  - $error_{true}(h) \geq \varepsilon$

Introduction to probability: outcomes

- An **outcome space** specifies the possible outcomes that we would like to reason about, e.g.

$\Omega = \{$  ,  $\}$ Coin toss

$\Omega = \{$  $\}$ Die toss

- We specify a **probability** p(**x**) for each outcome **x** such that

$$p(x) \geq 0, \qquad \sum_{x \in \Omega} p(x) = 1$$

E.g., p() = .6

p() = .4

Introduction to probability: events

- An **event** is a subset of the outcome space, e.g.

E = { , ,  }   Even die tosses

O = { , ,  }   Odd die tosses

- The **probability** of an event is given by the sum of the probabilities of the outcomes it contains,

$$p(E) = \sum_{x \in E} p(x)$$

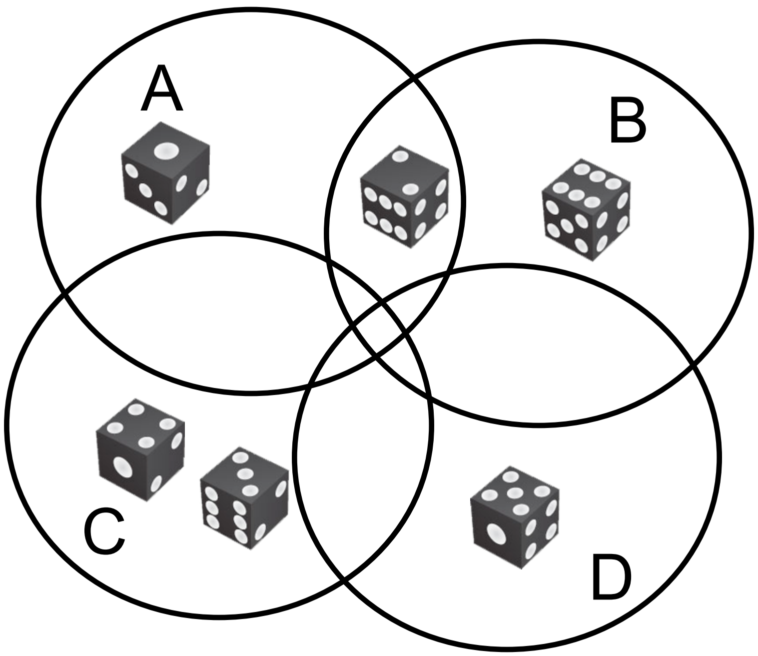
E.g., p(E) = p() + p() + p()

= 1/2, if fair die

Introduction to probability: union bound

- P(A or B or C or D or …)

  $\leq$ P(A) + P(B) + P(C) + P(D) + …



$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$
$$\leq p(A) + p(B)$$

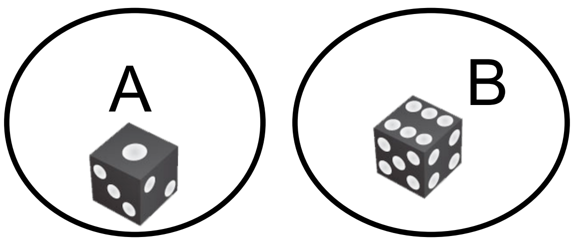
Q: When is this a tight bound?

A: For disjoint events
(i.e., non-overlapping circles)

Introduction to probability: independence

- Two events A and B are **independent** if

$$p(A \cap B) = p(A)p(B)$$



Are these events independent?

No! $p(A \cap B) = 0$

$$p(A)p(B) = \left(\frac{1}{6}\right)^2$$

Introduction to probability: independence

- Two events A and B are **independent** if

$$p(A \cap B) = p(A)p(B)$$

- Suppose our outcome space had two different die:

$\Omega = \{$  $\}$ 2 die tosses

$6^2 = 36$ outcomes

and the probability of each outcome is defined as

p() = $a_1\ b_1$ p() = $a_1\ b_2$ …

a_1	a_2	a_3	a_4	a_5	a_6
.1	.12	.18	.2	.1	.3

$$\sum_{i=1}^{6} a_i = 1$$

b_1	b_2	b_3	b_4	b_5	b_6
.19	.11	.1	.22	.18	.2

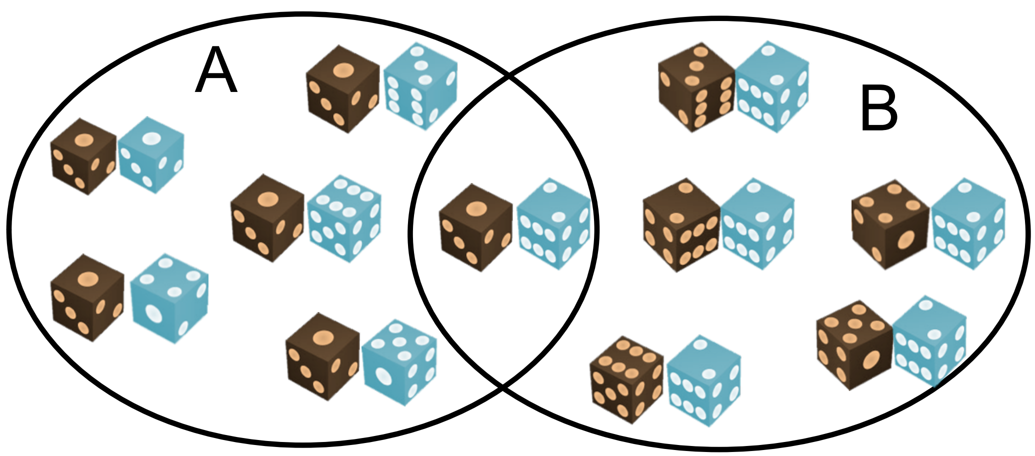
$$\sum_{j=1}^{6} b_j = 1$$

Introduction to probability: independence

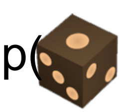
- Two events A and B are **independent** if

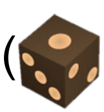
$$p(A \cap B) = p(A)p(B)$$

- Are these events independent?



Yes! $p(A \cap B) =$ p()

$p(A)p(B) =$ p() p()

p(A) = p()

$$= \sum_{j=1}^{6} a_1 b_j = a_1 \sum_{j=1}^{6} b_j = a_1$$

p(B) = p() $= b_2$