Machine Learning for Intelligent Systems

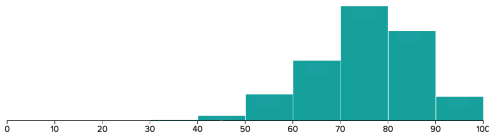
Lecture 17: Statistical Learning Theory 1

Reading: UML 4

Instructors: Nika Haghtalab (this time) and Thorsten Joachims

1

Prelim



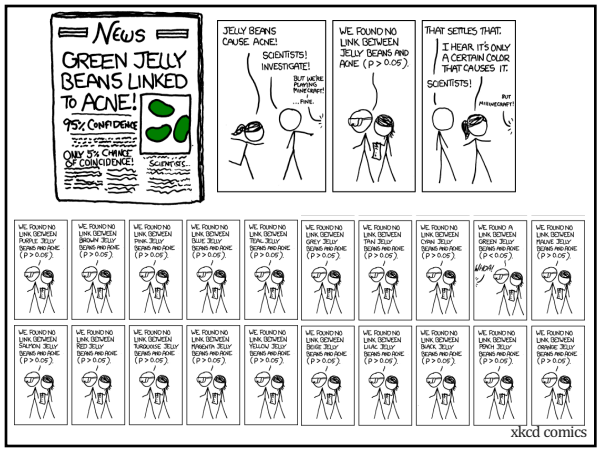
Curved:

$$CurvedGrade = 100 - 0.75(95 - RawPoints)$$

Harder time with the following concepts:

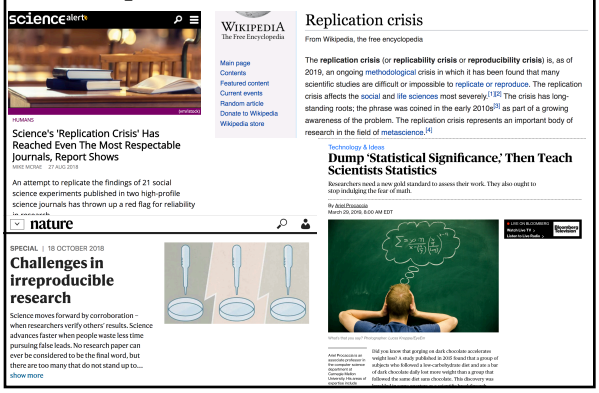
1. Perceptron update bound
2. Leave-on-out error of Kernelized SVM
3. Neural Network construction

2



3

Replication Crisis in Science



4

Convince me of your Psychic Abilities?

Game

- I'm thinking of m bits (0,1)
- If somebody in the class guesses my bit sequence, that person clearly has telepathic abilities – right?
- Think of a 6 digit 0,1 sequence.

Question:

- If at least one of $|H|$ players guesses the bit sequence correctly, is there any significant evidence that they have telepathic abilities?
- How large would m and $|H|$ have to be for us to trust this test?

5

Testing for psychic power

Set up:

- $|H|$ student $H = \{h_1, \dots, h_{|H|}\}$
- m bits (length of sequence)
- $p = 0.5$ probability of error on a single bit, if you're not psychic.

Prob. that student i guesses my code without being psychic?

$$P(h_i \text{ correct} \mid h_i \text{ not psychic}) = (1-p)^m$$

Prob. at least one student guesses my code, without anyone being psychic?

$$P(h_1 \text{ correct} \vee \cdots \vee h_{|H|} \text{ correct} \mid \text{nobody is psychic}) = 1 - (1-(1-p)^m)^{|H|}$$

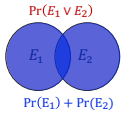
How long should the sequence be, so we are $1-\delta$ confident?

$$m > \log_{(1-p)}(1-(1-\delta)^{1/|H|})$$

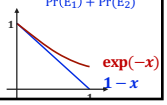
6

Useful Formulas

- **Binomial Distribution:** prob. of observing k heads in m independent coin tosses, where each toss is heads with prob. p, is
$$Pr(X = k|p, m) = \frac{m!}{k!\,(m-k)!} p^k (1-p)^{(m-k)}.$$
- **Hoeffding's inequality:** In the above binomial distribution,
$$\Pr\left[\left|\frac{k}{m} - p\right| > \epsilon\right] \leq 2\exp(-2m\epsilon^2)$$
- **Union Bound**: For any events E_i,
$$Pr(E_1 \vee E_2 \vee \cdots \vee E_k) \leq \sum_{i=1}^{k} Pr(E_i).$$



- **No name lemma:** $(1-\epsilon) \leq e^{-\epsilon}$



7

Fundamental Questions

Questions in Statistical Learning Theory:

- Trying to learn a classifier from H?
- How good is the learned rule after m examples?
- How many examples is needed for the learned rule to be accurate
- What can be learned and what cannot?
- Is there a universally best learning algorithm?

In particular, we will address:

- What kind of a guarantee on the true error of a classifier can I get if I know its training error?

8

Recall Prediction as Learning

9

Sample & Generalization Errors

Sample (Empirical) Error

Sample error of hypothesis h on samples $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, denoted by $err_S(h)$ is
$$err_S(h) = \frac{1}{m}\sum_{i=1}^{m} 1(h(x_i) \neq y_i)$$

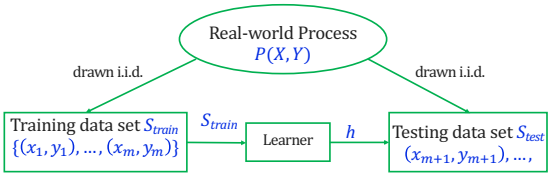
Generalization (Prediction/true) Error

Generalization error of hypothesis h on distribution $P(X,Y)$, denoted by $err_P(h)$ is
$$err_P(h) = \Pr_{(x,y)\sim P}[h(x) \neq y] = \sum_{i=1}^{m} 1(h(x) \neq y) \cdot P(X = x, Y = y)$$

10

Prediction as Learning

Goal: Find h with small prediction error $err_P(h)$ on $P(X,Y)$.
Strategy: Find an $h \in H$ with small sample error $err_{S_{train}}(h)$ on training dataset S_{train}.
Test the learned h to measure its **test error** $err_{S_{test}}(h)$ on a separate testing data set S_{test}.



11

Let's come back

12

Generalization Error Bounds

What kind of a guarantee on the true error of a classifier can I get if I know its training error?

Today's plan:

- **Zero empirical error:** If the rule I learned from H achieves zero error on the samples ($err_S(h) = 0$), how large is $err_P(h)$?
- **Non-zero empirical error:** How good is the true error of a hypothesis from H that that performs well on samples?

Today's assumption: The hypothesis set H is finite.

13

Zero Empirical Error

If the hypothesis I learned from H achieves zero error on the samples ($err_S(h) = 0$), how large is $err_P(h)$?

- **Assume** H is finite.
- **Assume Realizability:** There is a consistent classifier.

→There is always one $h \in H$ that $err_P(h) = 0$ – one person is psychic.

Algorithm $\mathcal{L}$ takes a set S of m samples from P and picks h_S that has 0 empirical error. What's the bound on $err_P(h)$?

1. Fix a hypothesis $h \in H$ before seeing S. What's the probability that $err_P(h) > \epsilon$, but $err_S(h) = 0$?
2. What's the probability that $err_P(h_S) > \epsilon$, but $err_S(h_S) = 0$?

14

Sample Complexity – 0 Empirical Error

Theorem

For any instance space X and set of labels $Y = \{-1, 1\}$ and for any distribution P on $X \times Y$, consider a set S of m i.i.d. samples from P, we have

$$\Pr_{S \sim P^m}[\exists\, h \in H, \text{such that } err_S(h) = 0, \text{but } err_P(h) > \epsilon] \leq |H| e^{-\epsilon m}.$$

Theorem: Sample Complexity (zero empirical error)

Let $m \geq \frac{1}{\epsilon}\left(\ln(|H|) + \ln\left(\frac{1}{\delta}\right)\right)$. For any instance space X, labels $Y = \{-1, 1\}$, distribution P on $X \times Y$, with probability $1 - \delta$ over i.i.d draws of set S of m samples, we have

Any $h \in H$ that has **0 empirical error**, has **true error** of $err_P(h) \leq \epsilon$.

Learning Algorithm: Given a sample set S and hypothesis class $h \in H$, if there is a $h_S \in H$ that is *consistent* with S, return h_S. (Eqv. Return h_S in version space $VS(H, S)$)

15