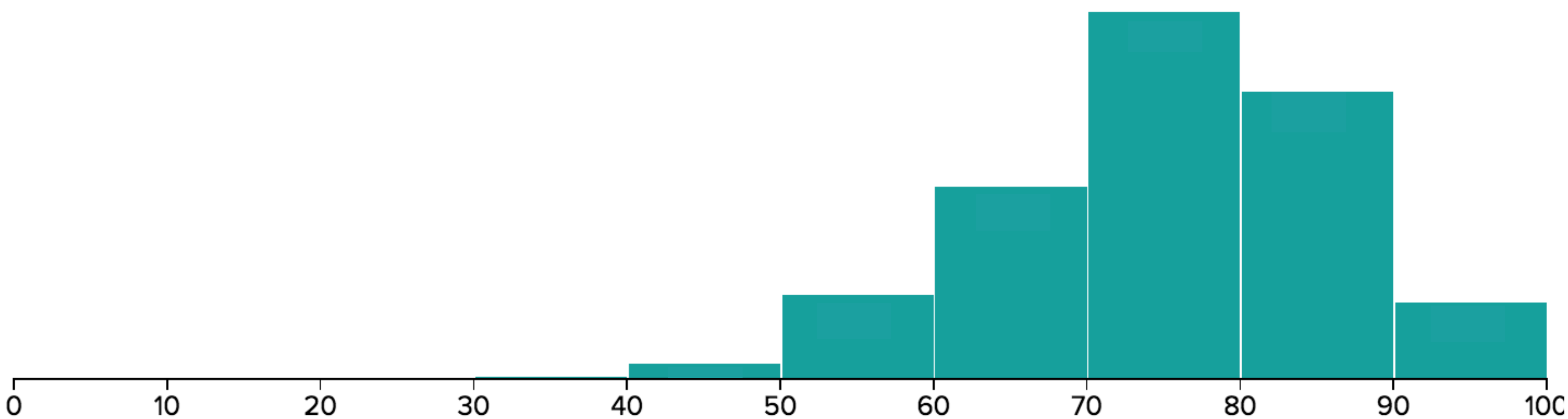
Machine Learning for Intelligent Systems

Lecture 17: Statistical Learning Theory 1

Reading: UML 4

Instructors: Nika Haghtalab (this time) and Thorsten Joachims

Prelim

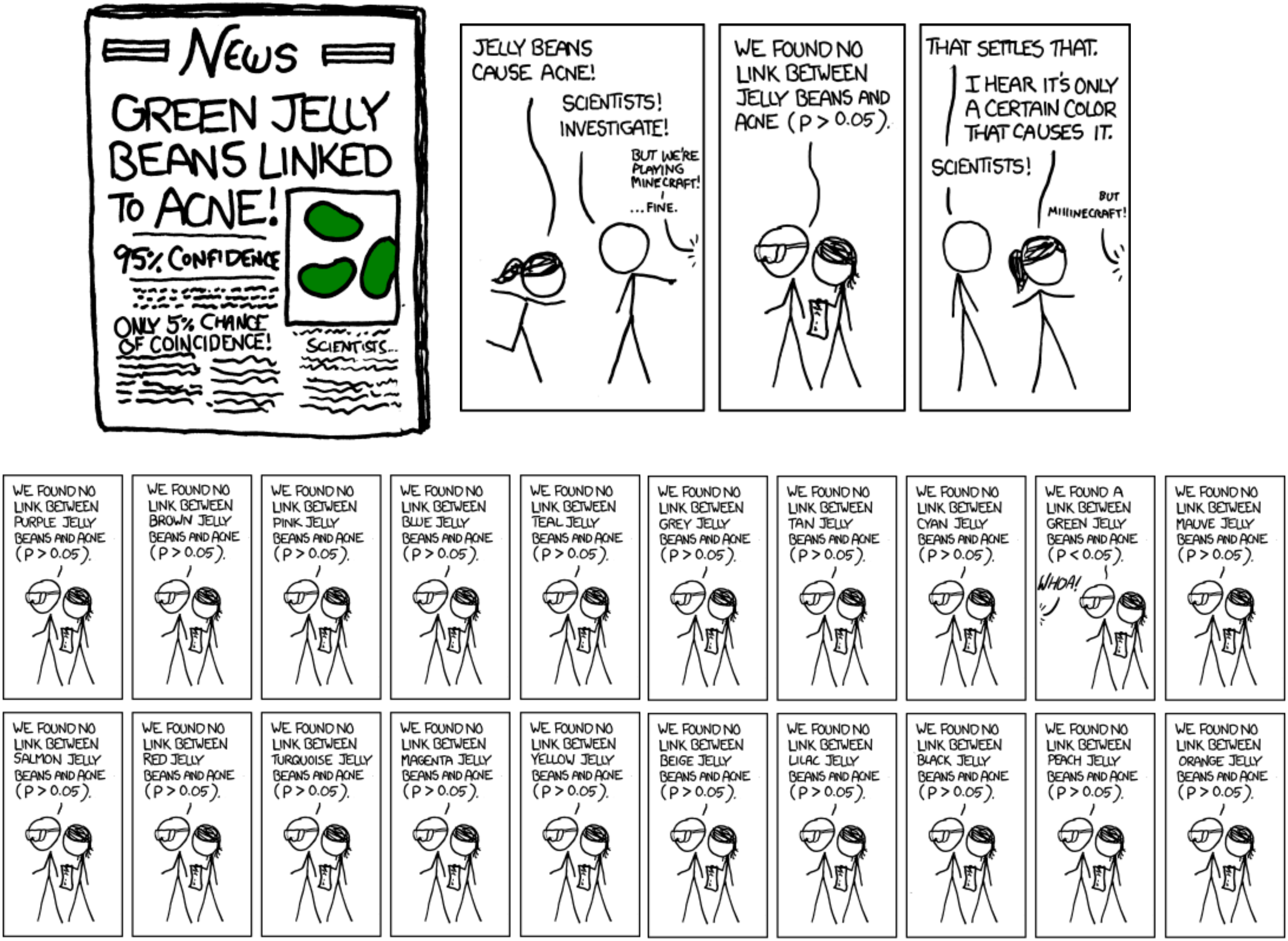


Curved:

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

Harder time with the following concepts:

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


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...
JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

xkcd comics

Replication Crisis in Science

sciencealert

(vm/istock)

HUMANS

Science's 'Replication Crisis' Has Reached Even The Most Respectable Journals, Report Shows

MIKE MCRAE 27 AUG 2018

An attempt to replicate the findings of 21 social science experiments published in two high-profile science journals has thrown up a red flag for reliability in research

Replication crisis

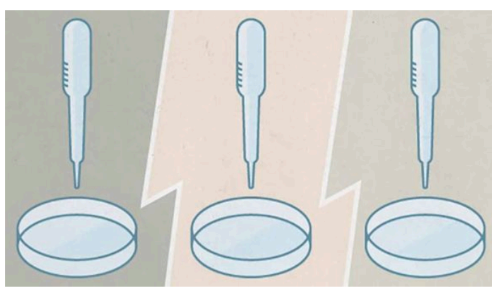
From Wikipedia, the free encyclopedia

The **replication crisis** (or **replicability crisis** or **reproducibility crisis**) is, as of 2019, an ongoing methodological crisis in which it has been found that many scientific studies are difficult or impossible to replicate or reproduce. The replication crisis affects the social and life sciences most severely.[1][2] The crisis has long-standing roots; the phrase was coined in the early 2010s[3] as part of a growing awareness of the problem. The replication crisis represents an important body of research in the field of metascience.[4]

nature

SPECIAL | 18 OCTOBER 2018

Challenges in irreproducible research



Science moves forward by corroboration – when researchers verify others' results. Science advances faster when people waste less time pursuing false leads. No research paper can ever be considered to be the final word, but there are too many that do not stand up to...

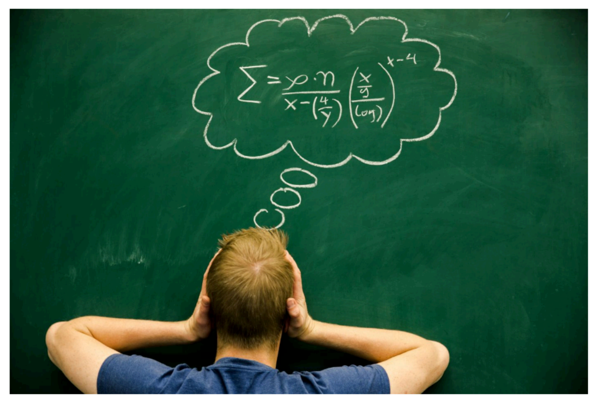
show more

Technology & Ideas

Dump 'Statistical Significance,' Then Teach Scientists Statistics

Researchers need a new gold standard to assess their work. They also ought to stop indulging the fear of math.

By Ariel Procaccia
March 29, 2019, 8:00 AM EDT



What's that you say? *Photographer: Lucas Knappe/EyeEm*

Ariel Procaccia is an associate professor in the computer science department at Carnegie Mellon University. His areas of expertise include artificial intelligence.

Did you know that gorging on dark chocolate accelerates weight loss? A study published in 2015 found that a group of subjects who followed a low-carbohydrate diet and ate a bar of dark chocolate daily lost more weight than a group that followed the same diet sans chocolate. This discovery was heralded in some quarters as a scientific breakthrough.

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?

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?

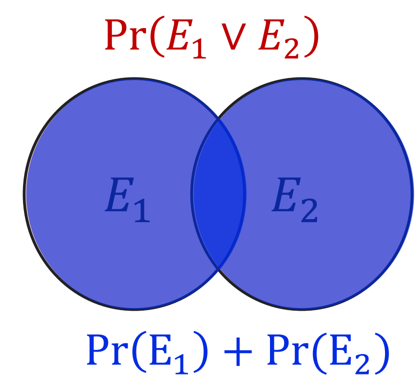
$$P\left(h_1 \text{ correct} \vee \cdots \vee h_{|H|} \text{ correct} \,\middle|\, \text{nobody is psychic}\right)$$
$$= 1 - (1 - (1-p)^m)^{|H|}$$

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

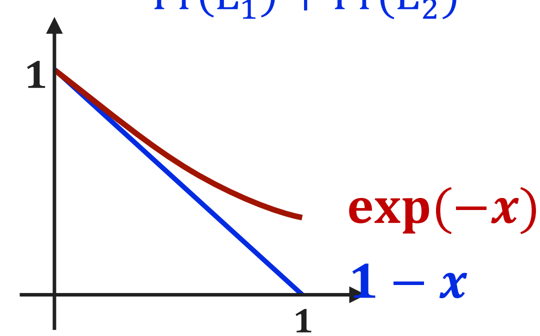
$$m > \log_{(1-p)}\left(1 - (1-\delta)^{1/|H|}\right)$$

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:** $(1-\epsilon) \leq e^{-\epsilon}$



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?

Recall Prediction as Learning

Sample & Generalization Errors

Sample (Empirical) Error

Sample error of hypothesis h on samples $\mathrm{S} = \{(x_1, y_1), \ \dots \ , (x_m, y_m)\}$, 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

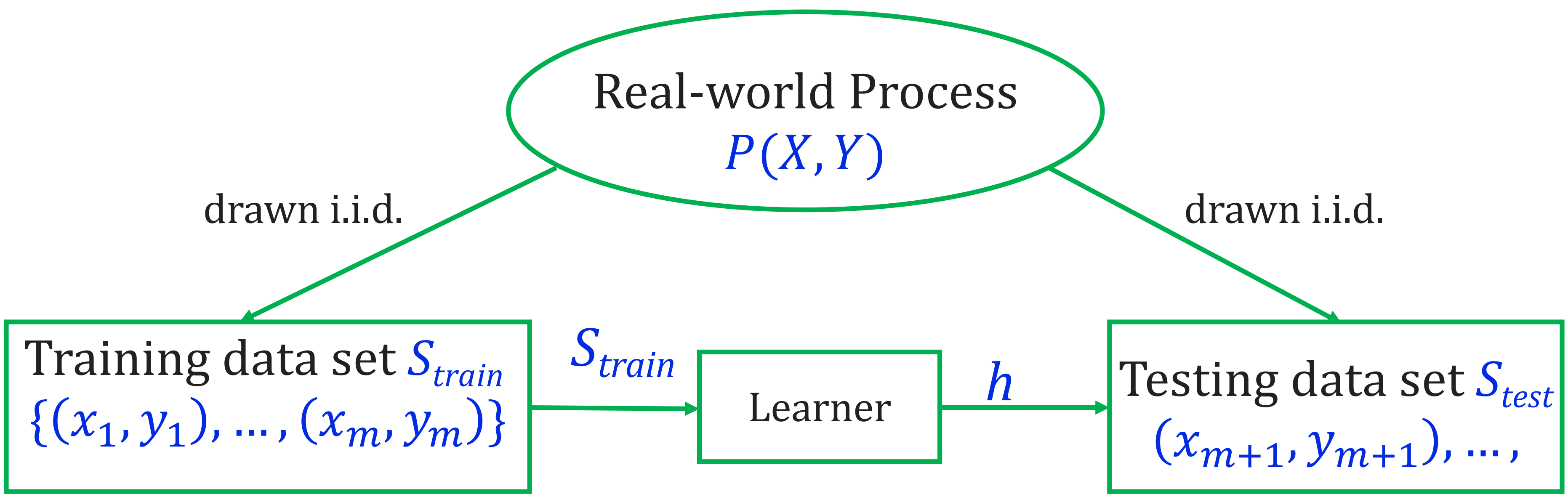
$$err_P(h) = \Pr_{(x,y) \sim P}[h(x) \neq y] = \sum_{i=1}^{m} 1(h(x) \neq y) \cdot P(X = x, Y = y)$$

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}.



Let’s come back

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.

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 $err_S(h_S) = 0$?

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

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

Theorem: Sample Complexity (zero empirical error)

Let $m \ge \frac{1}{\epsilon}\left(\ln(|H|) + \ln\left(\frac{1}{\delta}\right)\right)$. For any instance space X, labels $\mathrm{Y} = \{-1, 1\}$, 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 $err_P(h) \le \epsilon$.

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 $\mathrm{VS}(H, S)$)