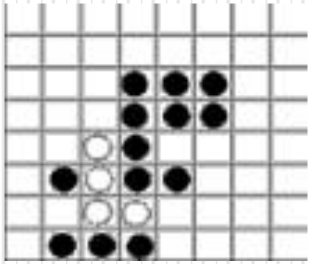
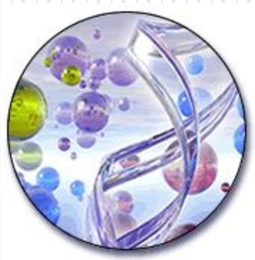
Welcome to

Introduction to Machine Learning!





Review of Topic 5: Hidden Markov Model

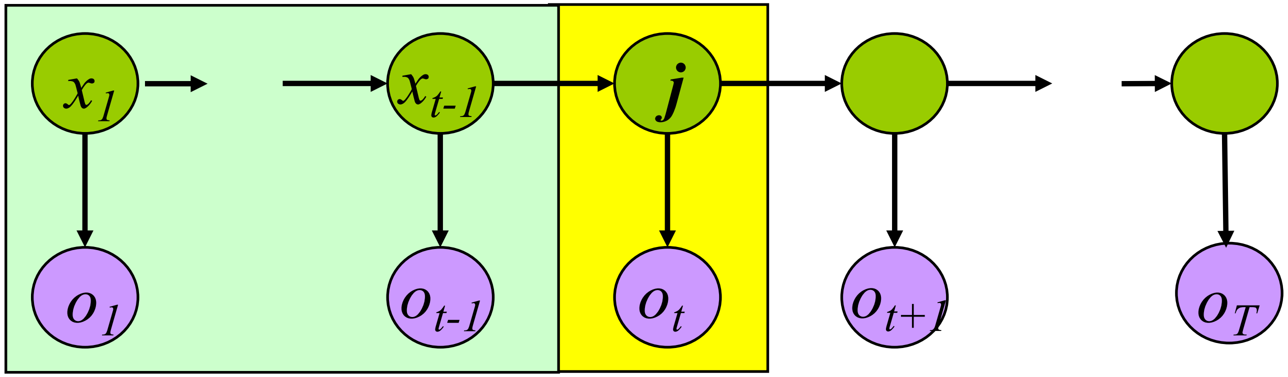
隐马尔可夫模型

Problem 2 – Solution 2

- Given an observation sequence, compute the most likely hidden state sequence
- Find the state sequence that best explains the observations
- There may be many X' s that make P(X|O) maximal.
- We give an algorithm to find one of them.

$$\arg\max_{X} P(X \mid O) \longrightarrow \textbf{Viterbi algorithm}$$

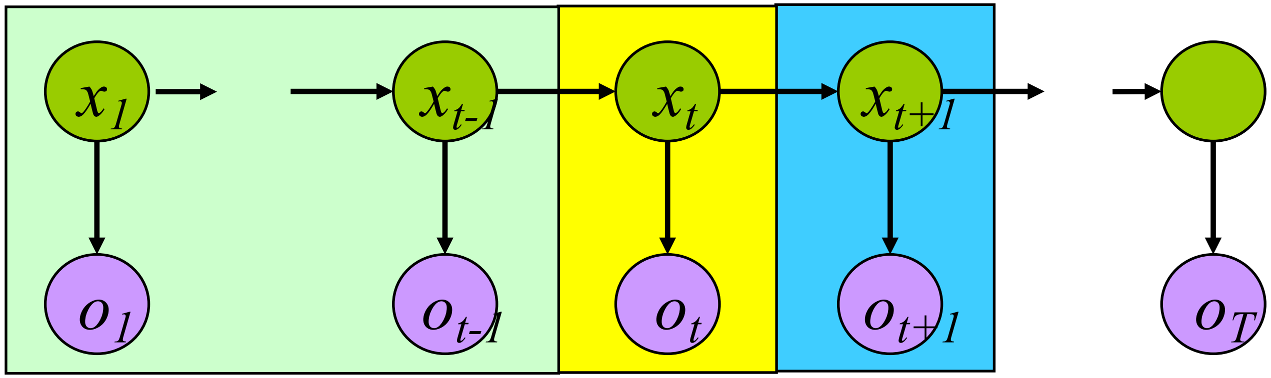
Viterbi Algorithm



$$\delta_t(j) = \max_{x_1 \ldots x_{t-1}} P(x_1 \ldots x_{t-1}, o_1 \ldots o_{t-1}, x_t = j, o_t)$$

The state sequence which maximizes the probability of seeing the observations to time t -1, landing in state j, and seeing the observation at time t

Viterbi Algorithm



$$\delta_t(j) = \max_{x_1 \ldots x_{t-1}} P(x_1 \ldots x_{t-1}, o_1 \ldots o_{t-1}, x_t = j, o_t)$$

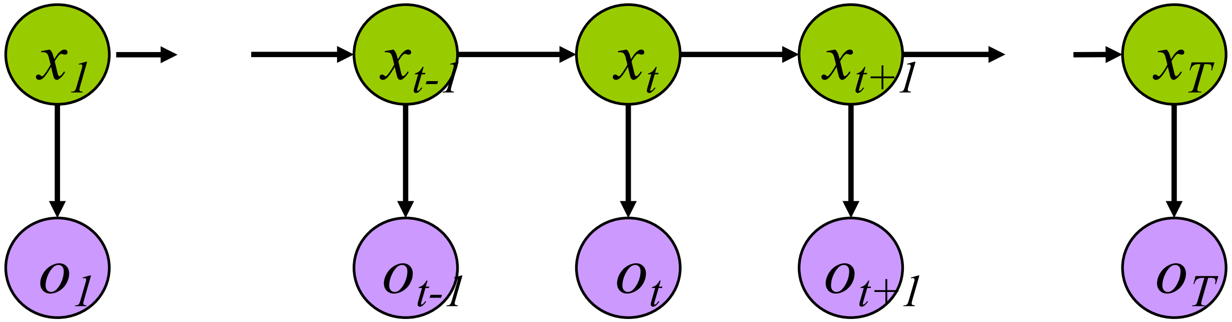
$$\delta_{t+1}(j) = \max_i \{\delta_t(i) a_{ij} b_{j o_{t+1}}\}$$

$$\psi_{t+1}(j) = \arg\max_i \{\delta_t(i) a_{ij} b_{j o_{t+1}}\}$$

Recursive

Computation

Viterbi Algorithm



$$\delta_t(j) = \max_{x_1 \ldots x_{t-1}} P(x_1 \ldots x_{t-1}, o_1 \ldots o_{t-1}, x_t = j, o_t)$$

$$\delta_{t+1}(j) = \max_i \{\delta_t(i) a_{ij} b_{j o_{t+1}}\} \quad \psi_{t+1}(j) = \arg\max_i \{\delta_t(i) a_{ij} b_{j o_{t+1}}\}$$

$$\hat{X}_T = \arg\max_i \delta_T(i)$$

$$P(\hat{X}) = \delta_T(i)$$

HMM applications

- Speech recognition
- (Chinese / Japanese) Input Method
- POS (Part of Speech) Tagging
- Gene analysis
- Any phenomena of linear sequence

Example: disease diagnose

- Observations

  Clinical symptoms c_i (fever, cough, sore throat, snivel, etc)

- States: diseases d_i (Influenza, pneumonia, tonsillitis, etc)
- Transition probabilities: $P(d_i \mid d_j)$
- Observation probabilities: $P(c_i \mid d_j)$
- Initialize state probabilities
- Encoding problem:
  - Clinical symptoms: cough → sore throat → snivel → fever
  - To find: what's the most probable disease changing sequence?

Example ：POS tagging (词性标注)

- Problem:

  Given word sequence $w_1 w_2 \ldots w_n$， find POS sequence $c_1 c_2 \ldots c_n$

- HMM model:
  - State: POS
  - Observation: Word
- Training:

  Learn POS transition matrix $[a_{ij}]$ and observation matrix (POS to words) $[b_{ik}]$ by statistical analyses

- Find the solution: Viterbi algorithm

Overview

- Estimation problem:
  - Define forward / backward variables
  - Dynamic algorithm, $O(N^2T)$
- Encoding problem: Viterbi algorithm
  - Dynamic algorithm, $O(N^2T)$

To master the definition and the computation of $\alpha_t(i)$

To master the viterbi algorithm

Overview (Cont.)

- Advantages of HMM:
  - Solid mathematical foundation, efficient algorithms, effective performance, easy to train
- The most import point:

We can use the special structure of this model to do a lot of neat math and solve problems that are otherwise not solvable.

- Further discussion:
  - Correctness and fitness of the 1st order Markov assumption

References

- L. Rabiner: A tutorial on Hidden Markov Models and selected applications in speech recognition, Proceedings of the IEEE 77(2): 257-286, 1989
  - Recommended readings: p257~p266 (exclude IV. Types of HMMs)
- E. Fosler-Lussier: Markov Models and Hidden Markov Models: A Brief tutorial, Technical Report (TR-98-041), December 1998, International Computer Science Institute, Berkeley, California.
- A. Meng: An introduction to Markov and Hidden Markov Models, http://eivind.imm.dtu.dk/teaching/04364/ex10/intro_HMM_01.pdf

Exercises: (by yourself)

Consider the following coin-tossing experiment:

	State 1	State 2	State 3
P(H)	0.5	0.75	0.25
P(T)	0.5	0.25	0.75

– state-transition probabilities equal to 1/3

– initial state probabilities equal to 1/3

1. You observe $O = (H, H, H, H, T, H, T, T, T, T)$. What state sequence, q, is most likely? What is the joint probability, $P(O, q|\lambda)$, of the observation sequence and the state sequence?

2. What is the probability that the observation sequence came entirely of state 1?

Topic 6. ML Theory-I: Evaluating Hypotheses
学习理论 I：假设的评估问题

Min Zhang
z-m@tsinghua.edu.cn

Review: Inductive learning hypothesis

- Much of the learning involves acquiring general concept from specific training examples.

- Inductive learning algorithms can at best guarantee that the output hypothesis fits the target concept over the training data.
  - **Notice: over-fitting problem**

Review: Inductive learning hypothesis

- The *Inductive Learning Hypothesis*:

  Any hypothesis found to approximate the target function well over **a sufficiently large set** of training examples will also approximate the target function well over unobserved examples.

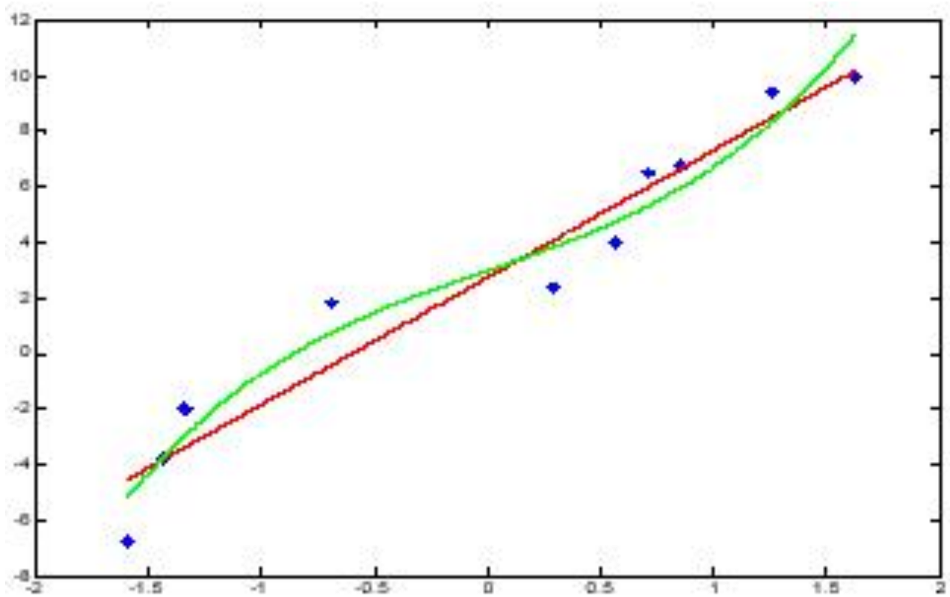
  (任一假设若在足够大的训练样例集中很好地逼近目标函数，它也能在未见实例中很好地逼近目标函数)

Motivation – Question 1

- Performance estimation
  - Given the observed accuracy of a hypothesis over a limited sample of data
  - how well does it estimate the accuracy over additional data?

Motivation – Question 2

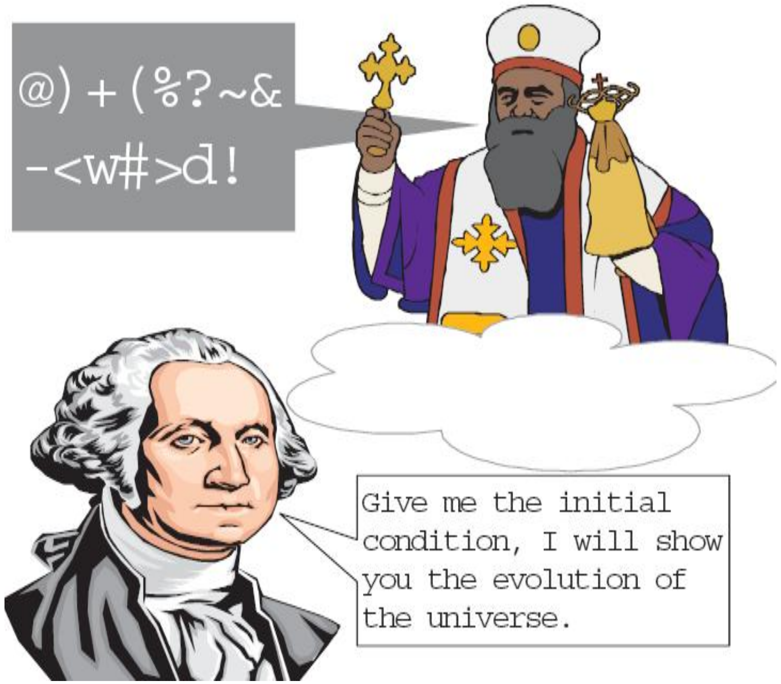
- h_1 outperforms h_2 over some sample of data
  - How probable is it that h_1 is better in general?



Motivation – Question 3

- When data is limited
  - what is the best way to use this data to both learn a hypothesis and estimate its accuracy?

The mathematical study of the likelihood and probability of events occurring based on known information and inferred by taking a limited number of samples.



Background Knowledge on Statistics

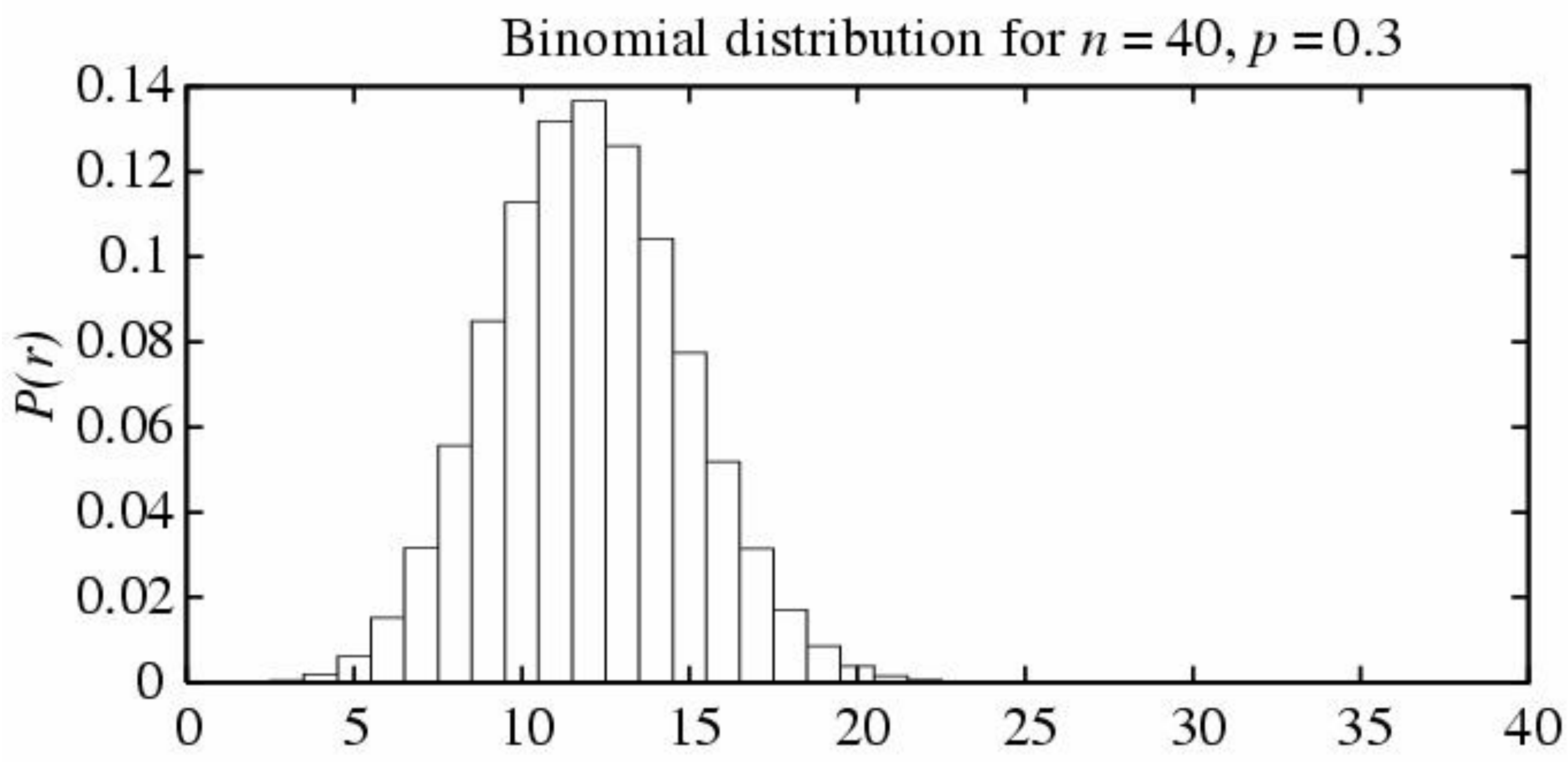
Basics of Sampling Theory

- Bernoulli experiments
  - Only 2 outputs: success probability: p, fail probability: $q = 1\text{-}p$
  - Use random variable X to record the number of success

- Binomial Distribution:
  - Toss a coin: probability of heads side up p, toss n times, observed heads up r times
  - If $X \sim B(n, p)$ then $Pr(X = r) = P(r)$

$$P(r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

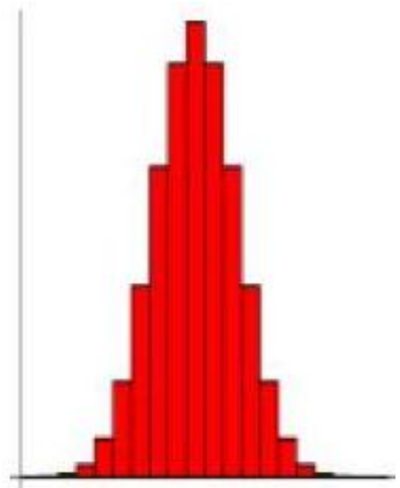
Binomial distribution



Where the Binomial Distribution Applies

- Two possible outcomes (success and failure) (Y=0 or Y=1)
- The probability of success is the same on each trial

  $Pr(Y=1)=p$, where p is a constant
- There are n independent trails
  - Random variables $Y_1,\ldots,Y_n$,
  - iid (independent identically distribution)
  - R: random variable, count of Y_i where $Y_i=1$ on n trails,
- $Pr(R=r) \sim$ Binomial distribution
- Mean (expected value): $E[R]$, μ
  - Binomial distribution: $\mu = np$
- Variance: Var[R]=E[(R-E[R])^2], σ^2 (Standard deviation σ)
  - Binomial distribution: $\sigma^2 = np(1-p)$



Discussions on Question 1

Review Question 1

- Performance estimation
  - Given the observed accuracy of a hypothesis over a limited sample of data
  - how well does it estimate the accuracy over additional data?

Estimating Hypothesis Accuracy: Define Problem

- Given:
  - A hypothesis *h* and a data sample containing *n* examples
  - Drawn at random according to the distribution D
  - Sample Error ***error$_S$*** $error_S(h) \equiv \frac{1}{n} \sum_{x \in S} \delta(f(x) \neq h(x))$
- Question
  1. What is the best estimate of the accuracy of *h* over future instances drawn from the same distribution?

     True Error ***error$_D$*** $error_{\mathcal{D}}(h) \equiv \Pr_{x \in \mathcal{D}}[f(x) \neq h(x)]$
  2. What is the probable error of the accuracy estimate?

Estimating Hypothesis Accuracy – answer to Q1.1

Back to Q1.1 What is the best estimate of the accuracy of h over future instances drawn from the same distribution?

Probability that r of n random samples are misclassified – Binomial distribution

$$P(r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

n=100, r =12 **n=25, r =3**

12% **12%**

$$error_D(h) = p \qquad error_s(h) = r/n$$

$$E[r] = np, \quad E[error_S(h)] = E[\frac{r}{n}] = \frac{np}{n} = p = error_D(h)$$

$$\sigma_{error_s(h)} = \frac{\sigma_r}{n} = \frac{\sqrt{np(1-p)}}{n} = \sqrt{\frac{error_D(h)(1-error_D(h))}{n}}$$

$$\sigma_{error_s(h)} = \frac{\sigma_r}{n} \approx \sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$$

3.2% **6.5%**

Two important properties of estimator

- Estimation bias
  - If S is training set, $error_S(h)$ is optimistically biased

$$bias \equiv E[error_S(h)] - error_D(h)$$

  - For unbiased estimate ($bias = 0$), h and S must be chosen independently → Don't test on training set!
- Estimation variance
  - Even with unbiased S, $error_S(h)$ may still vary from $error_D(h)$
    - E.g. previous examples of 3.2% vs. 6.5%
  - Should choose the unbiased estimator with least variance

Estimating Hypothesis Accuracy –Q1.2

- Q1.2 What is the probable error of the accuracy estimate?

(How well does $error_S(h)$ estimate $error_D(h)$?)

- Sampling theory: confidence interval (置信区间)
- Definition:
  - An N% confidence interval for some parameter p is an interval that is expected with probability N% to contain p.

  (N%: confidence degree)

  参数p 的N %置信区间是一个以N %的概率包含p 的区间，

  N% : 置信度

Confidence interval

- Example of confidence interval
  - Suppose you know nothing about a girl, therefore her age is a random variable X for you.
  - When you see her photo, you guess her age between 12 to 18 with confidence level 90%
  - To get more confidence level (e.g. 99.9%), the interval has to be larger (e.g. [3,50])
- How to get confidence interval?
  - Bad news: Hard with Binomial Distribution
  - Good news: Easy with Normal Distribution
    - Obtained with area (integral) of normal distribution

Normal distribution

- Probability density function of normal dist.

Normal Dist. & Binomial Dist.

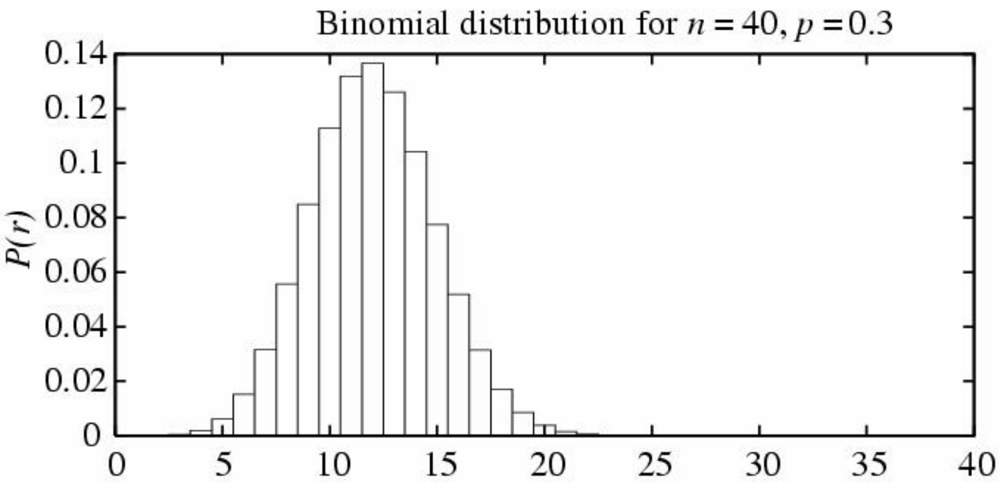
- For sufficiently large sample sizes

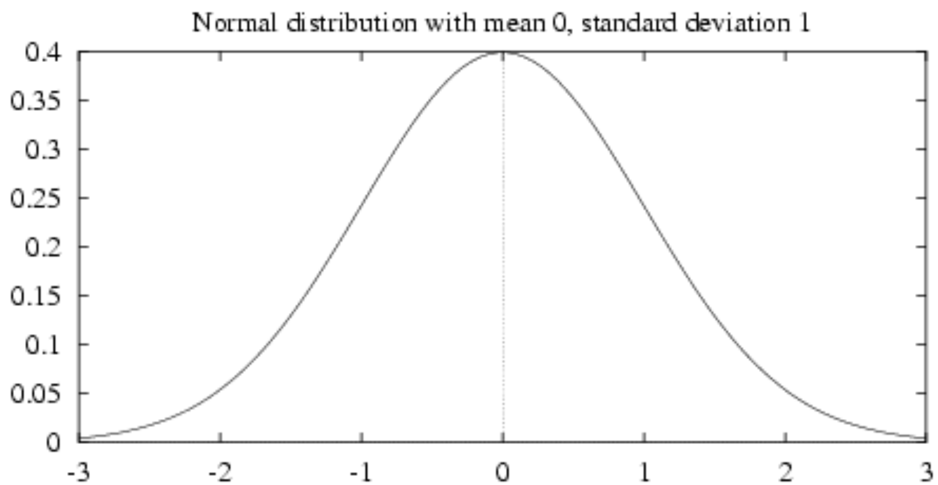
The binomial distribution

can be closely approximated by

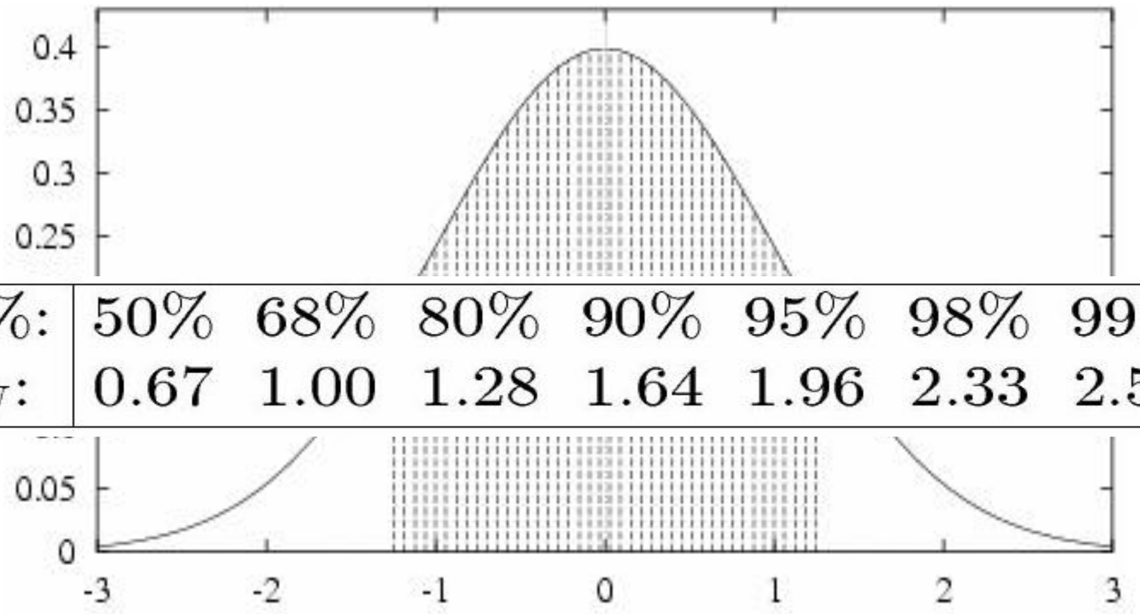
The Normal distribution

- Rule of thumb: n>30, $np(1\text{-}p)$> 5





Confidence Interval of Normal Distribution



$N\%$:	50%	68%	80%	90%	95%	98%	99%
z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58

- 80% of area (the measured value y) lies in $\mu \pm 1.28\sigma$
- N % of area (the measured value y) lies in $\mu \pm z_N \sigma$
- Equivalently, the mean μ will fall in the following interval $N\%$ of the time

  $y \pm z_N \sigma$

Estimating Hypothesis Accuracy – the answer to Q1.2

- More correctly, if
  - S contains n examples, drawn independently of h and each other, n >= 30
- Then
  - With approximately 95% probability, $error_S(h)$ lies in interval

$$error_{\mathcal{D}}(h) \pm 1.96\sqrt{\frac{error_{\mathcal{D}}(h)(1 - error_{\mathcal{D}}(h))}{n}}$$

Estimating Hypothesis Accuracy – the answer to Q1.2

- More correctly, if
  - S contains n examples, drawn independently of h and each other, $n >= 30$
- Then
  - With approximately 95% probability, $error_S(h)$ lies in interval

$$error_{\mathcal{D}}(h) \pm 1.96\sqrt{\frac{error_{\mathcal{D}}(h)(1 - error_{\mathcal{D}}(h))}{n}}$$

equivalently $error_{\mathcal{D}}(h)$ lies in interval

$$error_S(h) \pm 1.96\sqrt{\frac{error_{\mathcal{D}}(h)(1 - error_{\mathcal{D}}(h))}{n}}$$

Estimating Hypothesis Accuracy – the answer to Q1.2

- More correctly, if
  - S contains n examples, drawn independently of h and each other, $n >= 30$
- Then
  - With approximately 95% probability, $error_S(h)$ lies in interval

$$error_{\mathcal{D}}(h) \pm 1.96\sqrt{\frac{error_{\mathcal{D}}(h)(1 - error_{\mathcal{D}}(h))}{n}}$$

equivalently $error_{\mathcal{D}}(h)$ lies in interval

$$error_S(h) \pm 1.96\sqrt{\frac{error_{\mathcal{D}}(h)(1 - error_{\mathcal{D}}(h))}{n}}$$

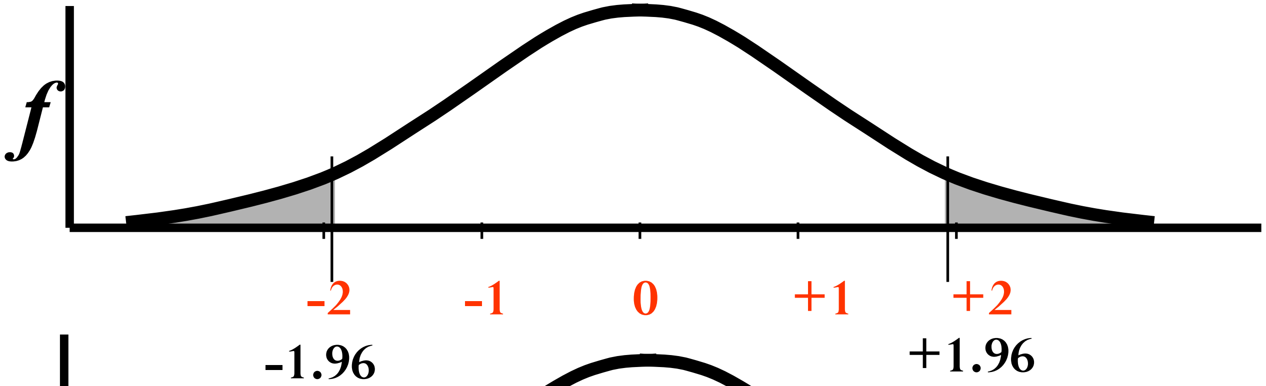
which is approximately

$$error_S(h) \pm 1.96\sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

More details: One-sided bounds

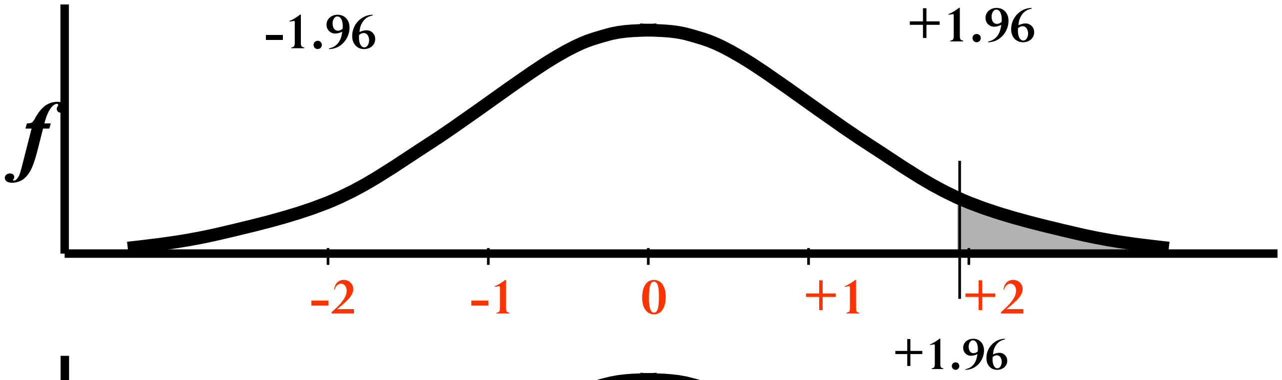
- Yield upper or lower error bounds
- We know the probability that $error_D(h)$ lies in $[L,U]$
- Then what's the probability that $error_D(h)$ is less than U ?
  - Symmetry of Normal Distribution

More details: One-sided bounds



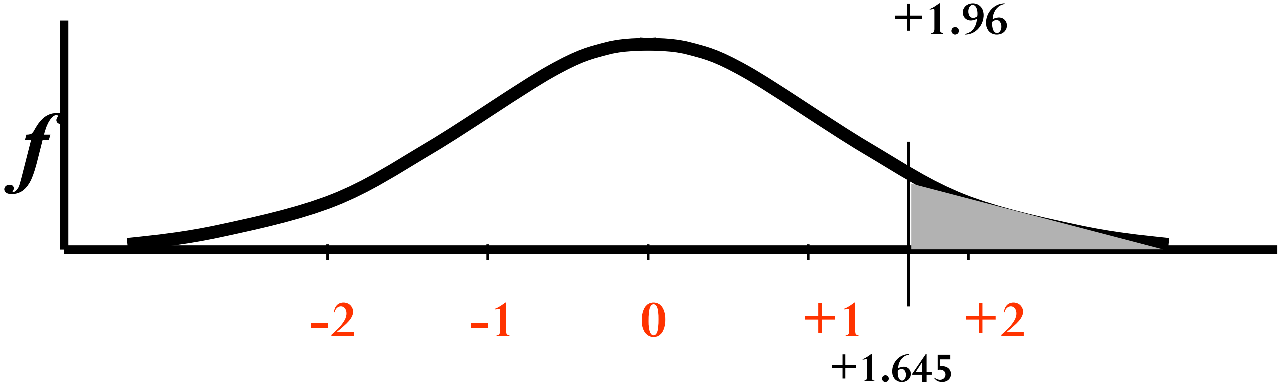
$z_{CV} = \pm 1.96$

$N\% = 95\%$



$z_{CV} = + 1.96$

$N\% = 97.5\%$

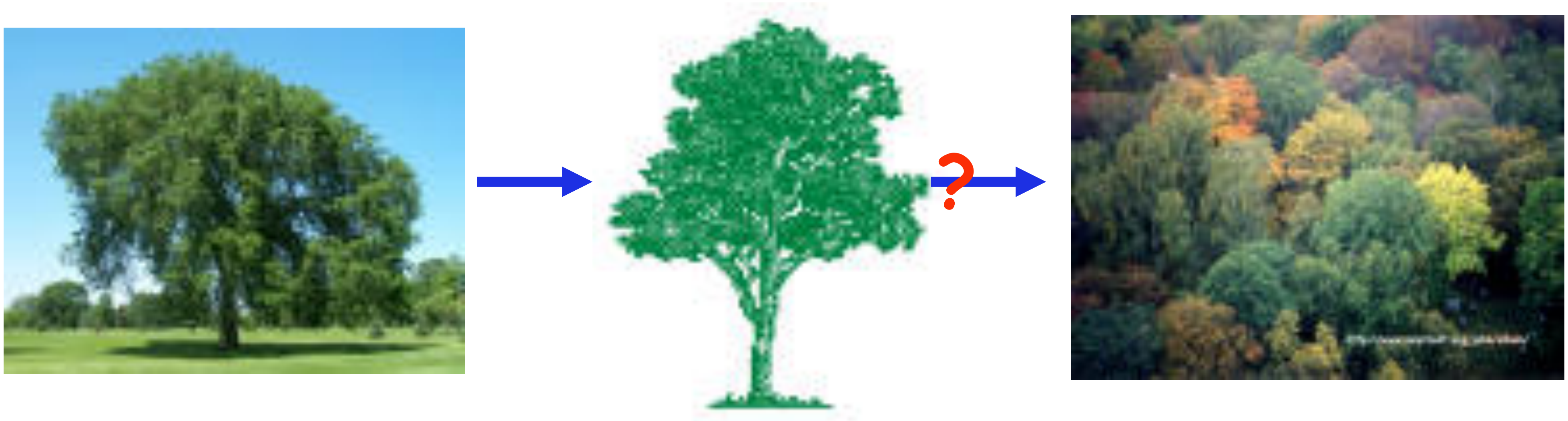


$z_{CV} = + 1.645$

$N\% = 95\%$

Recall – Question 1

- Performance estimation
  - Given the observed accuracy of a hypothesis over a limited sample of data
  - how well does this estimate its accuracy over additional data?



Overview: Answers to Question 1

- Problem setting:
  - S: n random independent samples, and independent with hypothesis h
  - $n >= 30$ & h with r errors
- True error $error_D$ lies in the following interval with N% confidence:

$$error_S(h) \pm z_N \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

$N\%$:	50%	68%	80%	90%	95%	98%	99%
z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58

More Information on

Deriving Confidence Intervals

General Approach for deriving Confidence Intervals

- In general
  - Identify the parameter p to estimate, e.g. $error_D(h)$
  - Define estimator Y (bias, variance), e.g. $error_S(h)$
    - Desirable : minimum variance, unbiased estimator
  - Determine distribution D governing Y (including mean & variance)
  - Determine N% confidence interval $(L..U)$
    - Could have L=-∞ or U=∞
    - E.g. Use table of z_n values (for normal distribution)
- Applied later to other problems

General Approach for deriving Confidence Intervals

- In general
  - Identify the parameter p to estimate, e.g. $error_D(h)$
  - Define estimator Y (bias, variance), e.g. $error_S(h)$
    - Desirable : minimum variance, unbiased estimator
  - Determine distribution D governing Y (including mean & variance)
  - Determine N% confidence interval $(L..U)$
    - Could have $L=-\infty$ or $U=\infty$
    - E.g. Use table of z_n values (for normal distribution)
- Applied later to other problems

Central Limit Theorem

- Simplifies attempts to define confidence intervals.
- Problem setting
  - Independent, identically distributed (iid) random variable $Y_1, .. , Y_n$,
  - unknown distribution, with mean μ and finite variance σ^2
  - Estimating mean: $\bar{Y} \equiv \frac{1}{n} \sum_{i=1}^{n} Y_i$
- Central Limit Theorem
  - $\bar{Y}$ approaches a normal distribution ($n \rightarrow \infty$)
  - With mean μ , and variance σ^2/n
  - Can be normalized to the normal dist. with $\mu = 0$, $\sigma = 1$

Central Limit Theorem …

- Distribution of sample mean $\vec{Y}$
  - is known
  - although distribution of Y_i is not
  - can be used to determine mean & variance of Y_i
- Gives basis to approximating
  - Distribution of estimators
  - That are means of some sample

Application: DTree Avoid over-fitting

- Two ways of avoid over-fitting for D-Tree

  I. Stop growing when data split not statistically significant (pre-pruning)

  II. Grow full tree, then post-pruning

For option II:

- How to select "best" tree?
  - Measure performance over training data (statistical pruning)
    - Confidence level
  - Measure performance over separate validation data set

Decision Tree Pruning based on Confidence Intervals (as in C4.5)

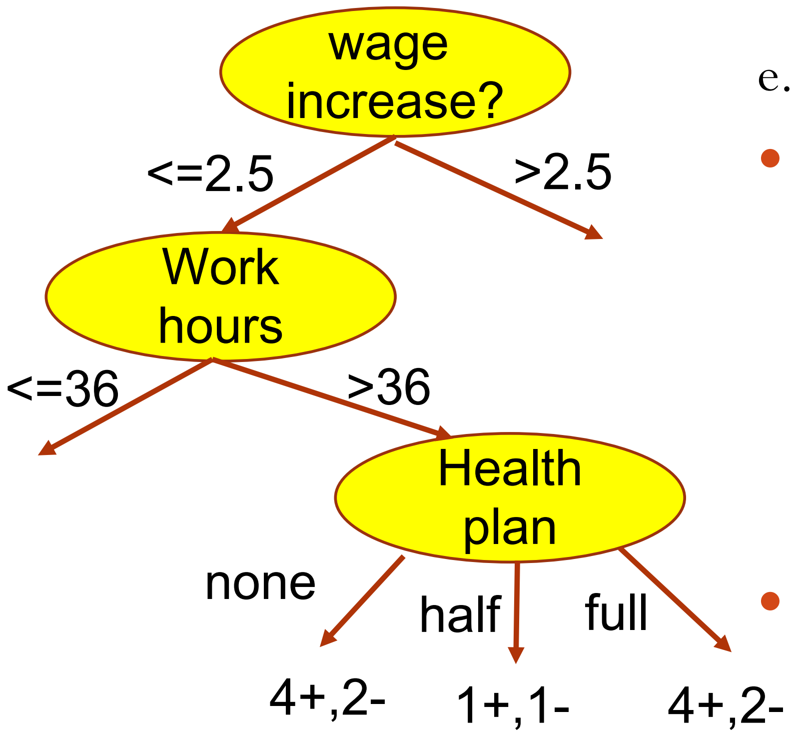
- Advantage: It allows all of the available labeled data to be used for training.
- Key idea: calculate a confidence interval for the error rate.
- True error $error_D$ lies in the following interval with N% confidence:

$$error_S(h) \pm z_N \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

$N\%$:	50%	68%	80%	90%	95%	98%	99%
z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58

- In order to decide whether to replace a near-leaf node and its child leaves by a single leaf node, C4.5 compares the upper limits of the error confidence intervals for the two trees
- For the unpruned tree, the upper error estimate is calculated as a weighted average over its child leaves.

Decision Tree Pruning based on Confidence Intervals (as in C4.5)



e.g. For health plan node: (75% confidence, z=0.69)

- The average estimated upper error rate for the unpruned tree
  - =none: err_s=2/6, n=6, err_t upper bound: 0.46
  - =half: err_s=1/2, n=2, err_t upper bound: 0.74
  - =full: err_s=2/6, n=6, err_t upper bound: 0.46
  - Weighted average upper error rate: **0.50**
- If the node "health plan" is pruned → leaf(9+,5-)
  - err_s=5/14, n=14, Estimated err_t upper bound: **0.44**
- The pruned tree results in a lower upper estimate for the error rate, the leaves are indeed pruned.

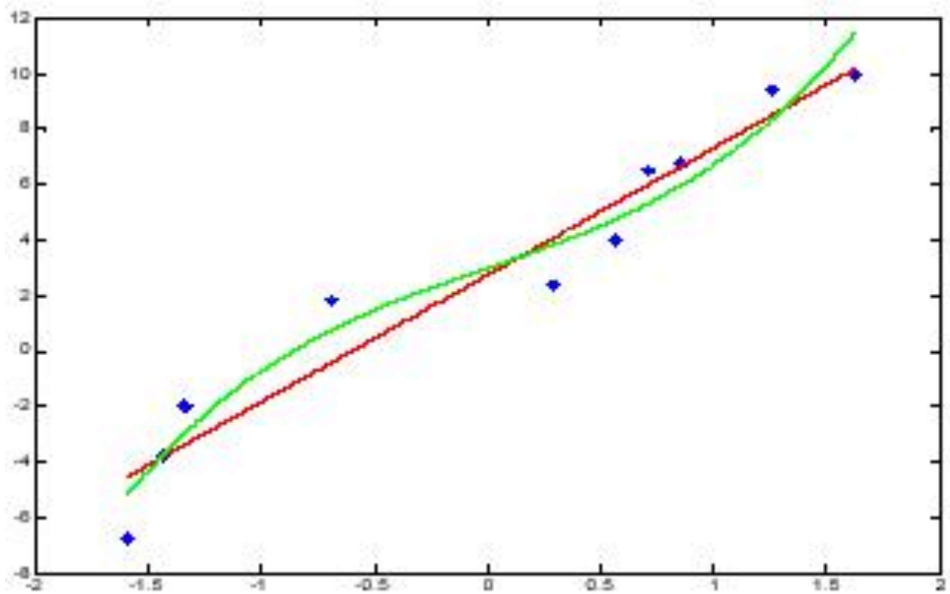
More details and progress: Gilad Katz, Asaf Shabtai, Lior Rokach, and Nir Ofek. Confdtree: Improving decision trees using confidence intervals. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 339 –348, dec. 2012

Discussions on Question 2

Question 2

- h_1 outperforms h_2 over some sample of data
  - How probable is it that h_1 is more accurate in general?

Difference between hypotheses



Difference between hypotheses

- Test h_1 on sample S_1 (n_1 random samples), test h_2 on S_2 (n_2)
- Pick parameter to estimate $d \equiv error_{\mathcal{D}}(h_1) - error_{\mathcal{D}}(h_2)$
- Choose an estimator $\hat{d} \equiv error_{S_1}(h_1) - error_{S_2}(h_2)$
  - Unbiased
- Determine probability distribution that governs estimator
  - $error_{S1}(h_1)$, $error_{S2}(h_2)$ approx. Normal Dist.
  - $\hat{d}$ is also approx. Normal Dist. *
    - Mean = d
    - variances: sum up

* Proof: http://en.wikipedia.org/wiki/Sum_of_normally_distributed_random_variables

Difference between hypotheses

- Test h_1 on sample S_1 (n_1 random samples), test h_2 on S_2 (n_2)
- Pick parameter to estimate $d \equiv error_{\mathcal{D}}(h_1) - error_{\mathcal{D}}(h_2)$
- Choose an estimator $\hat{d} \equiv error_{S_1}(h_1) - error_{S_2}(h_2)$
  - Unbiased
- Determine probability distribution that governs estimator

$$\sigma_{\hat{d}} \approx \sqrt{\frac{error_{S_1}(h_1)(1 - error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1 - error_{S_2}(h_2))}{n_2}}$$

Difference between hypotheses

- Test h_1 on sample S_1 (n_1 random samples), test h_2 on S_2 (n_2)
- Pick parameter to estimate $d \equiv error_{\mathcal{D}}(h_1) - error_{\mathcal{D}}(h_2)$
- Choose an estimator $\hat{d} \equiv error_{S_1}(h_1) - error_{S_2}(h_2)$
  - Unbiased
- Determine probability distribution that governs estimator

$$\sigma_{\hat{d}} \approx \sqrt{\frac{error_{S_1}(h_1)(1 - error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1 - error_{S_2}(h_2))}{n_2}}$$

- Find interval (L,U) such that N % of probability mass falls in the interval

$$\hat{d} \pm z_N \sqrt{\frac{error_{S_1}(h_1)(1 - error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1 - error_{S_2}(h_2))}{n_2}}$$

Hypothesis testing

- Probability that some claim is true
  - E.g. prob. that $e_D(h_1) > e_D(h_2)$
- Example (n_1=n_2=100)
  - $e_{S1}(h_1) = 0.3$, $e_{S2}(h_2) = 0.2$, prob. that $e_D(h_1) > e_D(h_2)$

$$\hat{d} \equiv error_{S_1}(h_1) - error_{S_2}(h_2)$$

- Given $\hat{d}$ =0.1, prob. that $e_D(h_1) > e_D(h_2)$

Hypothesis testing

- Probability that some claim is true
  - E.g. prob. that $e_D(h_1) > e_D(h_2)$
- Example (n_1=n_2=100)
  - $e_{S1}(h_1) = 0.3$, $e_{S2}(h_2) = 0.2$
    - Given $\hat{d}$ =0.1, prob. that $e_D(h_1) > e_D(h_2)$
    - Given $\hat{d}$ =0.1, prob. that $d > 0$

$$d \equiv error_{\mathcal{D}}(h_1) - error_{\mathcal{D}}(h_2)$$

$$\hat{d} \equiv error_{S_1}(h_1) - error_{S_2}(h_2)$$

Hypothesis testing

- Probability that some claim is true
  - E.g. prob. that $e_D(h_1) > e_D(h_2)$
- Example (n_1=n_2=100)
  - $e_{S1}(h_1) = 0.3$, $e_{S2}(h_2) = 0.2$
    - Given $\hat{d}$ =0.1, prob. that $e_D(h_1) > e_D(h_2)$
    - Given $\hat{d}$ =0.1, prob. that $d > 0$
    - Prob. $\hat{d}$ is in interval $d + 0.1 > \hat{d}$
      - Note: d is the mean of distribution of $\hat{d}$
    - Prob. $\hat{d}$ is in interval $\hat{d} < \mu_{\hat{d}} + 0.1$

$$d \equiv error_{\mathcal{D}}(h_1) - error_{\mathcal{D}}(h_2)$$

$$\hat{d} \equiv error_{S_1}(h_1) - error_{S_2}(h_2)$$

$$\mu \pm z_N \sigma$$

Hypothesis testing (cont.)

$$e_{S1}(h_1) = 0.3,\ e_{S2}(h_2) = 0.2 \quad n_1 = n_2 = 100$$

- Approx. distribution of $\hat{d}$ is known

$$\sigma_{\hat{d}} \approx \sqrt{\frac{error_{s_1}(h_1)(1-error_{s_1}(h_1))}{n_1} + \frac{error_{s_2}(h_2)(1-error_{s_2}(h_2))}{n_2}} = 0.061$$

$$\hat{d} < \mu_{\hat{d}} + 0.1 \rightarrow \hat{d} < \mu_{\hat{d}} + 1.64\sigma_{\hat{d}}$$

- $Z_N = 1.64$, 90% two-sided confidence interval (c.i.)
- i.e. 95% one-sided c.i.
- $e_D(h_1) > e_D(h_2)$ with 95% confidence

Hypothesis testing (cont.)

$$e_{S1}(h_1) = 0.3,\ e_{S2}(h_2) = 0.2 \qquad n_1 = n_2 = 30$$

- Approx. distribution of $\hat{d}$ is known

$$\sigma_{\hat{d}} \approx \sqrt{\frac{error_{s_1}(h_1)(1-error_{s_1}(h_1))}{n_1} + \frac{error_{s_2}(h_2)(1-error_{s_2}(h_2))}{n_2}} = \mathbf{0.111}$$

$$\hat{d} < \mu_{\hat{d}} + 0.1 \rightarrow \hat{d} < \mu_{\hat{d}} + \mathbf{0.90}\ \sigma_{\hat{d}}$$

- $Z_N =$ **0.90** , **68%** two-sided confidence interval (c.i.)
- i.e. **84%** one-sided c.i.
- $e_D(h_1) > e_D(h_2)$ with **84%** confidence

Discussions on Question 3

Question 3

- When data is limited
  - What is the best way to use this data to both learn a hypothesis and estimate its accuracy?

Comparing learning algorithm

- We would like to estimate:

$$E_{S \subset \mathcal{D}}[error_{\mathcal{D}}(L_A(S)) - error_{\mathcal{D}}(L_B(S))]$$

  - where L(S) is the hypothesis output by learner L using training set S.
  - "Which one is better on average?"
  - Performance over all S drawn from D, independent test set
- But, given limited data D_0, what is a good estimator?
  - Divide D_0 into training set S_0 and testing set T_0 and measure:

$$error_{T_0}(L_A(S_0)) - error_{T_0}(L_B(S_0))$$

Holdout

  - Even Better, repeat this many times and average the results.
  - Paired t-test: 2 algorithms use same training and testing sets

Comparing learning algorithm

1. Partition data D_0 into k disjoint test sets $T_1, T_2, \ldots, T_k$ of equal size, where this size is at least 30.
2. For i from 1 to k, do

   use T_i for the test set, and the remaining data for training set S_i

   - $S_i \leftarrow \{D_0 - T_i\}$
   - $h_A \leftarrow L_A(S_i)$
   - $h_B \leftarrow L_B(S_i)$
   - $\delta_i \leftarrow error_{T_i}(h_A) - error_{T_i}(h_B)$
3. Return the value $\bar{\delta}$, where

$$\bar{\delta} \equiv \frac{1}{k} \sum_{i=1}^{k} \delta_i$$

***k* -fold cross validation**

Confidence intervals

- z_N can not be used
  - The training sets in this algorithm are not independent. (**they overlap!**)

$$|S| = \frac{k-1}{k} |D_0|$$

- $t_{N,k}$ & estimate of deviation

$$\bar{\delta} \pm t_{N,k-1} \; s_{\bar{\delta}} \qquad s_{\bar{\delta}} \equiv \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (\delta_i - \bar{\delta})^2}$$

- k = *degrees of freedom*
  - # of independent random events affecting the variable value $\bar{\delta}$
- Confidence interval (c.i.) of Paired t-test: a tighter c.i.
  - **Any differences in observed errors in a paired test are due to differences between the hypotheses**.

Practical setting

- This method is not strictly "statistically valid"
  - Re-sampling → training sets not independent → δ_i not independent
  - Testing set is independent
  - But even this approximation is better than no comparison.
- Other sampling techniques
  - Draw random test sets ($|S| > 30$)
    - Drawback: test sets may overlap

Overview：Answers to the 3 questions

1. Estimating hypothesis accuracy, confidence

Binomial Dist.→ Normal Dist., Confidence interval

2. h_1 outperforms h_2 over some samples

- In general, h_1 is better than h_2 ?

Difference of hypotheses → to find one-sided c.i.

3. How to use limited data to learn and estimate?

Paired *t*-test, *k*-fold cross validation, c.i. with $t_{N,k-1}$

For more info...

- More references
  - Dietterich, T. G., (1998). "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms." *Neural Computation, 10* (7) 1895-1924
  - Kong EB., Dietterich TG., "Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree Algorithms." technical report. (1995).
- **Recommended reading**
  - **正态分布的前世今生(上) http://cos.name/2013/01/story-of-normal-distribution-1/**
  - **正态分布的前世今生(下) http://cos.name/2013/01/story-of-normal-distribution-2/**

Homework

(1) Tom Mitchell, Machine learning, Exercise 5.4 (p152, En.)

(2) Evaluate your NB classifiers in Experiment 1

- Compare results on 5% & 100% training set respectively
  - I. Estimations of $error_D$ and the C.I. respectively
  - II. What's the confidence of algorithm A is better than B in general?

- Submit deadline: **March 30 (Thursday 11:59pm).**