Information Retrieval

WS 2015 / 2016

Lecture 12, Tuesday January 26th, 2016
(Hypothesis Testing, Statistical Significance)

Prof. Dr. Hannah Bast
Chair of Algorithms and Data Structures
Department of Computer Science
University of Freiburg

Overview of this lecture

- Organizational
  - Your experiences with ES11 Perceptrons
  - The official **evaluation** of this course
- Contents
  - Perceptron refinements recap + two new ones
  - Hypothesis testing motivation + terminology
  - Randomization test example + program
  - Z-Test and T-Test example + math behind
  - **Exercise Sheet 13:** improve basic Perceptron algorithm + check whether the improvement is statistically significant

Experiences with ES11 1/2

- Summary / excerpts
  - Nice + interesting exercise again
  - The proof was easy / nice / doable / ok
  - Numpy still annoying, but getting used to it
  - Problems with operations **between** numpy and scipy
  - Lack of time due to other courses and deadlines
  - Linear algebra rulez … **YES !**

Experiences with ES11 2/2

- Results
  - Precision is comparable to that of Naïve Bayes

    Comedy vs. Thriller: Perceptron 87% NB 85%

    R vs. Non-R: Perceptron 70% NB 74%
  - The training is much slower than for Naïve Bayes

    It can be made much faster using "batching" … see slide 10
  - The top words are more meaningful than with Naïve Bayes

    Comedy vs. Thriller: comedy, thriller, noir, suspense, …

    R vs. Non-R: pg, spielberg, sex, slasher, neo, …

Official course evaluation 1/2

- Instructions
  - You should have received an email from EvaSys Admin on Monday, January 25 with a link to an evaluation form
  - We are **very** interested in your feedback
  - Please take your time for this
  - Please be honest and concrete
  - The **free text comments** are most interesting for us

  Please complete by Tuesday, February 9

  The evaluation is centralized, and will be closed after that date, and there is nothing we can do about that

Official course evaluation 2/2

- Why you should invest the time
  - If you have done the exercise sheets:

    Compared to the effort for the sheets, the evaluation is a piece of cake … take it
  - If you have not done the exercise sheets:

    If we receive much less feedback than in the last years, exercise sheets will be mandatory again next year
  - If you have neither did the exercise sheets nor attended the lectures nor listened to the recordings:

    Well … good luck with the exam

Perceptron Refinements 1/4

- Refinements we already discussed
  - Change the (pre-determined) number of iterations
  - Terminate when change in precision (on training set) drops below a certain threshold
  - Remove frequent words
  - Use tf.idf instead of tf to represent documents
  - Use different / additional features, e.g. word bigrams

Perceptron Refinements 2/4

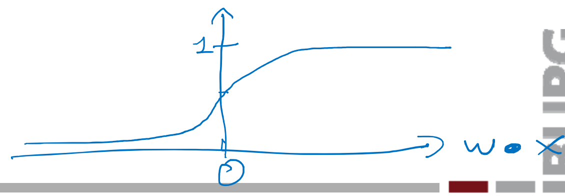
- Averaging
  - Take the average of all w from all iterations … including all the iterations where w did not change

    That is, if you have 10 iterations and a training set of size 100, you take the average of 1000 w vectors
  - Intuition 1: the final changes to w are due to relative few documents (which are still misclassified)

    Averaging de-emphasizes the w vectors from the end
  - Intuition 2: good values of w are not changed for many iterations (where they classify elements correctly)

    Averaging emphasizes those "good" w vectors

Perceptron Refinements 3/4



sigmoid function

w.l.o.g b=0 (incorporated into w) see L11

- Logistic Regression
  - Let $S(t) = 1 / (1 + e^{-t})$ ... then $S(w \bullet x)$ can be interpreted as the probability that x is classified as $+1$
  - We can now try to find the w such that the observed data is most likely ... another instance of MLE
  - This gives the following refined update step:

    Class of x is $+1$: $w \leftarrow w + \alpha \cdot a \cdot x$

    Class of x is -1 : $w \leftarrow w - \alpha \cdot a' \cdot x$

    where $a = 1 - S(w \bullet x)$ and $a' = S(w \bullet x)$ and α is a tuning parameter (the so-called learning rate)

Perceptron Refinements 4/4

- Batching
  - Given a w, consider a whole batch B of training elements

    The size of the batch is a parameter to play around with
  - For each $x_i \in B$ compute update term with respect to w

    Simple Perceptron: $+ x$ if class is $+1$, $-x$ otherwise

    "Logistic" Perceptron: $+ \alpha \cdot a \cdot x$ … or … $- \alpha \cdot a' \cdot x$
  - Then add all the update terms to w to obtain a new w

    Batching mainly improves performance (a lot), but it also affects the precision (since it leads to a different w)

Hypothesis Testing 1/6

- Motivation
  - Typical situation in research: compare the outcome of two experiments

    E.g. in the **life sciences**: health status for two groups of people, one taking a particular medication and one not

    E.g. in **computer science**: the performance of two systems, using different algorithms or different parameter settings
  - The outcome of the experiments will be different

    But even carrying out the same experiment twice will give different results because of random fluctuations

    Key question: how to tell a "real" difference between the two experiments from mere random fluctuation

Hypothesis Testing 2/6

- Example 1: Prediction of coin tosses
  - Ten predictions in a row, C = correct, W = wrong

    CCCCCCCCCC (all ten predictions are correct)
  - Do we believe in this person's ability to predict?
  - Let's assume H_0 = the person cannot predict, that is, is just making random guesses … with $Pr(C) = ½$

    H_0 is called the null hypothesis … see slide 14
  - Then $Pr(\text{all ten correct} \mid H_0) = 2^{-10} \le 0.001 = 0.1\%$

    Very unlikely that this great prediction was mere chance

Hypothesis Testing 3/6

$\binom{10}{8} = \frac{10 \cdot 9}{1 \cdot 2} = 45$

$\binom{10}{9} = \frac{10}{1} = 10$

$\binom{10}{10} = 1$

- Example 2: Prediction of coin tosses
  - Let us now assume a slightly less stellar prediction:

    CCCWCCCWCC (8 correct, 2 wrong)

  - What is now the probability that this is due to chance?

    Note: we should **not** ask for the probability of **exactly 8** correct guesses to happen; it makes more sense to ask for the prob. of **8 or more** correct guesses to happen

  - $\Pr(\geq 8 \text{ correct} \mid H_0) = \binom{10}{8} \cdot 2^{-10} + \binom{10}{9} \cdot 2^{-10} + \binom{10}{10} \cdot 2^{-10}$

    $= 56 \cdot 2^{-10} \approx 5\%$

Hypothesis Testing 4/6

- General approach
  - Hypothesis H e.g. ability to predict coin tosses
  - Null hypothesis H_0 e.g. random guessing (the opposite of H)
  - Compute the probability p of the given <u>or more extreme</u> data assuming that H_0 is true

    This probability p is called the **p-value**
  - If p is small enough, the observations are said to be **statistically significant** with significance level p

    In the life sciences, people are usually happy with values of $p < 0.05$ (moderate significance) $p < 0.01$ (strong sign.)

Hypothesis Testing 5/6

- Example 3: two dice with unknown distribution
  - Two dice A and B, four rolls each

    A : 1 , 3 , 3 , 5

    B : 6 , 6 , 4 , 4
  - Null hypothesis H_0 = the two dice A and B are identical
  - Given H_0, what is the probability of observing A and B

    This will be our running example for the rest of todays lecture

Hypothesis Testing 6/6

- Well known hypothesis tests
  - R-Test: simple + makes no probabilistic assumptions
  - Z-Test: assume normal distribution with fixed variance
  - T-Test: like Z-test, but also model variance distribution

R(andomization)-Test 1/3

- One of the simplest statistical tests
  - Assume we have two series of measurements, A and B
  - Null hypothesis = no difference between A and B
  - Then we can assume that the measurements come from one experiment + assignment to either A or B is arbitrary
  - The R-Test considers all 2^n possible assignments of the n measurements to either A or B
  - For each assignment, compute the difference $\Delta\mu$ of the means, and see if it is $\geq$ the $\Delta\mu$ on the observed data

    The fraction of assignments for which this is the case is the p-value according to the R-Test

R(andomization)-Test 2/3

- Application to our dice example

  A : 1 , 3 , 3 , 5 $\quad \mu_1 = \frac{1+3+3+5}{4} = 3$

  B : 6 , 6 , 4 , 4 $\quad \mu_2 = \frac{6+6+4+4}{4} = 5$ $\Rightarrow \Delta\mu = 2$

  – Here are some of the 2^8 possible assignments of these 8 measurements to either A or B and the respective Δμ

  Note: we ignore the two assignments, where all measurements are assigned all to A or all to B, because we can't compute a meaningful mean difference then

	1	3	3	5	6	6	4	4	
≥ 2	A	A	A	A	B	B	B	B	$\mu_1 = 3,\ \mu_2 = 5,\ \Delta\mu = 2$
< 0	B	B	B	B	B	B	A	A	$\mu_1 = 4,\ \mu_2 = 4,\ \Delta\mu = 0$

...

R(andomization)-Test 3/3

- Continuation of the example
  - Let's write a program together to iterate over all $2^8 - 2$ assignments and compute the p-value as explained

    254
  - Observation: for 46 of the assignments, the difference of the means is 2 or larger → $p = 46 / 254 \approx 18.1\%$
  - Note: for a small number n of measurements, we can easily try out (on a computer) all $2^n - 2$ assignments

    But for larger n, this quickly becomes infeasible

    For $n = 30$ we already have $2^{30} \approx 1$ billion assignments

    Then we can take a (large enough) random sample of assignments and compute the fraction for those

Z-Test and T-Test 1/12

- Assumptions
  - The Z-Test and the T-Test both assume an underlying probability distribution
  - Z-Test: underlying **normal distribution**
  - T-Test: underlying **t-distribution**
  - Then, for our setting, the p-value is $\Pr(M \geq \Delta\mu)$, where:

    M is a random variable, modelling the difference of the means with the assumed probability distribution

    $\Delta\mu$ is the value of M on the observed measurements

    As a preparation, let us recap (on the next slides) some foundations from probability theory …

Z-Test and T-Test 2/12

$E(X-EX)^2 = E(X^2 - 2X\,EX + (EX)^2)$
$= EX^2 - 2\,EX \cdot EX + (EX)^2$
$= EX^2 - 2(EX)^2 + (EX)^2$
$= EX^2 - (EX)^2$

- Random variables
  - Continuous random variable X = range is **R**
  - Cumulative distribution function: $\Phi(x) = \Pr(X \leq x)$

    In particular: $\lim_{x\to\infty} \Phi(x) = 1$
  - Mean: $\mathbf{E}\,X := \int (1 - \Phi(x))\,dx$

    In the discrete case, $\mathbf{E}\,X = \sum_k \Pr(X \geq k)$
  - Variance: $\mathbf{var}(X) := \mathbf{E}\,(X - \mathbf{E}\,X)^2 = \mathbf{E}\,X^2 - (\mathbf{E}\,X)^2$

    The square root of the variance is often called **standard deviation**, and often denoted by σ … then $\mathrm{var}(X) = \sigma^2$

Z-Test and T-Test 3/12

- Basic linearity properties of **E** and **var** :
  - For all X, Y it holds that: $\mathbf{E}(X + Y) = \mathbf{E}X + \mathbf{E}Y$

    Surprising but true: even if X and Y are dependent
  - For X, Y independent: $\mathbf{var}(X + Y) = \mathbf{var}(X) + \mathbf{var}(Y)$

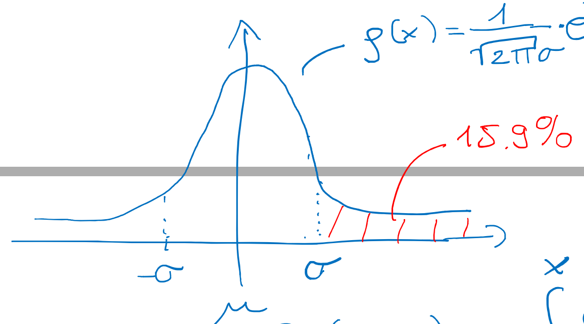
    Not generally true when X and Y are dependent
  - For X and any real a : $\mathbf{var}(a \cdot X) = a^2 \cdot \mathbf{var}(X)$

    This can be easily proved from the definition of var

$$\mathrm{var}(a \cdot X) = E(a \cdot X)^2 - (E\, a \cdot X)^2$$
$$= a^2 \cdot EX^2 - a^2 (EX)^2 = a^2 \cdot \mathrm{var}(X) \quad \boxtimes$$

Z-Test and T-Test 4/12

$$\varrho(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



$$Pr(X \le x) = \int_{-\infty}^{x} \varrho(t)\,dt$$

- The normal distribution
  - Assumed as the underlying distribution in many scenarios

    In the life sciences as well as in computer science
  - Two parameters: the mean μ and the variance σ^2

    The corresponding distribution is denoted by $N(\mu, \sigma^2)$
  - We will need to compute $Pr(X \geq x)$ where X has normal dist.

    There is no closed formula for this … in the ancient past, lookup tables were used

    For ES12, use scipy.stats.norm.cdf to obtain $Pr(X \leq x)$

Z-Test and T-Test 5/12

standard normal distribution

- Properties of the normal distribution
  - **Property 1:** If X has distribution $N(\mu, \sigma^2)$, then $(X - \mu) / \sigma$ has distribution $N(0, 1)$

    Every normal distr. can be reduced to N(0, 1) by scaling
  - **Property 2:** If X_1 has distribution $N(\mu_1, \sigma_1^2)$ and X_2 has distribution $N(\mu_2, \sigma_2^2)$, and X_1 and X_2 are independent then $X_1 + X_2$ has distribution $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

    The sum of normal random variables is again normal

Z-Test and T-Test 6/12

- Properties of the normal distribution, continued
  - **Property 3:** Let $X_1,\ldots,X_n$ be n **i.i.d.** (independent identically distributed) random variables, each with mean μ and variance σ^2 . Then $(X_1 + \ldots + X_n) / n$ converges to $N(\mu, \sigma^2)$ as $n \to \infty$

    This is known as the Central Limit Theorem

    It is the reason why the normal distribution is a natural assumption for many quantities observed in the world

    (for example, think of the running time of a loop with n iterations, and X_i = the time for the i-th iteration)

$X_i \sim N(\mu, \sigma^2)$

$\Rightarrow \frac{X_i - \mu}{\sigma} \sim N(0,1)$

Z-Test and T-Test 7/12

- The χ^2 **distribution** χ = small Greek letter "chi"
  - Let $Z_1, \ldots, Z_n$ be i.i.d. from $N(0, 1)$
  - Then the distribution of $Z = Z_1^2 + \ldots + Z_n^2$ is defined as: the χ^2 distribution with n degrees of freedom aka $\chi^2(n)$
  - Consider measurements $X_1, \ldots, X_n$, each from $N(\mu, \sigma^2)$

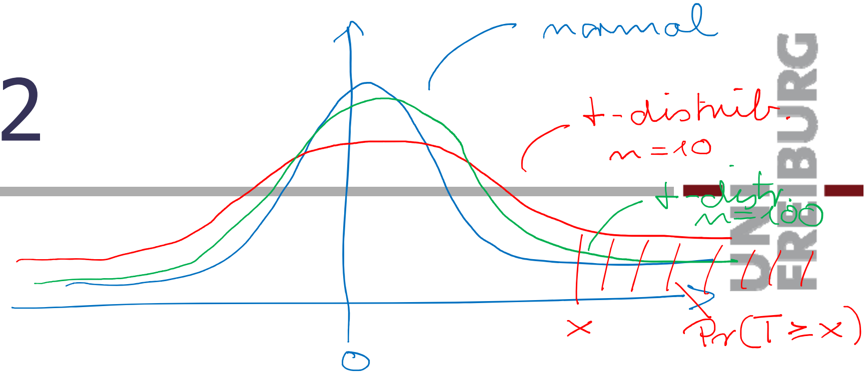
    Let $M = \Sigma X_i / n$ be the estimated mean, $\mathbf{E}\, M = \mu$

    Let $S^2 = \Sigma (X_i - M)^2 / n$ be the estimated variance, $\mathbf{E}\, S^2 = \sigma^2$

    Then $S^2 \cdot n / \sigma^2 = \Sigma ((X_i - M) / \sigma)^2$ has a $\chi^2(n)$ distribution

    Intuitively: the variance of a series of measurements has a χ2 distribution (up to scaling)

Z-Test and T-Test 8/12



- Student's t-distribution
  - Let us define it by how we pick a random X from it, in comparison to the standard normal distribution:

    Standard normal distribution: pick X from N(0, 1)

    T-distribution with n d.o.f: pick V from $\chi^2(n)$, then pick X from N(0, n / V)

  - Note that **E** V = n (slide 26) and that for $n \to \infty$ we have $V / n \to 1$ and the two distributions become the same

    Actually, there is a marked difference between the two distributions only for small n, say n ≤ 50

    For ES12, use scipy.stats.t.cdf to obtain Pr(X ≤ x)

Z-Test and T-Test 9/12

- More intuition about the difference
  - By also considering the variance as a random variable, the t-distribution is **less concentrated** around its mean than the corresponding normal distribution
  - Here is an example which provides some intuition

    Experiment 1: pick X uniformly from [-10, 10]

    Experiment 2: first pick V uniformly from [5, 15], then pick X uniformly from [-V, V]

    Now extreme values (< -10 or > 10) become more likely, and values around the mean become less likely

    Note that the mean remains zero in Experiment 2

Z-Test and T-Test 10/12

- The Z-Test assumption: underlying **normal distribution**
  - Given two series X_1 and X_2 of a total of n measurements
  - Let $M = M_1 - M_2$ be the difference of the means of X_1 and X_2
  - Let $S^2 = (\Sigma (X_{1j} - M_1)^2 + \Sigma (X_{2j} - M_2)^2) / (n/2)$ be the est. var.
  - Let $\Delta\mu$ and σ be the observed value of M and S, respectively
  - **H_0:** all $X_{ij} \sim N(\mu, \sigma^2)$

    Naïve assumption: the real variance is the observed variance
  - Then $Z = \sqrt{n} \cdot M / (2\sigma)$ has distribution $N(0, 1)$
  - The p-value of the Z-Test is then $\Pr(M \geq \Delta\mu) = \Pr(Z \geq x)$ where $x = \sqrt{n} \cdot \Delta\mu / (2\sigma)$

Z-Test and T-Test 11/12

- The T-Test assumption: underlying **t-distribution**
  - Given two series X_1 and X_2 of a total of n measurements
  - Let $M = M_1 - M_2$ be the difference of the means of X_1 and X_2
  - Let $S^2 = (\Sigma (X_{1j} - M_1)^2 + \Sigma (X_{2j} - M_2)^2) / (n/2)$ be the est. var.
  - Let $\Delta\mu$ and σ be the observed value of M and S, respectively
  - **H_0:** all $X_{ij} \sim N(\mu, S^2)$, $S^2 \sim \sigma^2 / (V/n)$, $V \sim \chi^2(n)$ with n d.o.f.

    More realistic: the underlying variance is a random variable
  - Then $T = \sqrt{n} \cdot M / (2S)$ has t-distribution with n d.o.f.
  - The p-value of the T-Test is then $\Pr(M \geq \Delta\mu) = \Pr(T \geq x)$, where $x = \sqrt{n} \cdot \Delta\mu / (2\sigma)$

Z-Test and T-Test 10a/12

- The Z-Test improved slide, for use in future course
  - Given two series X_1 and X_2 of a total of n measurements

    Common unknown mean $\mu = \mathbf{E}\, X_{ij}$ and variance $\sigma = \mathbf{var}\, X_{ij}$
  - Let $M = M_1 - M_2$ be the difference of the means of X_1 and X_2
  - Let $S^2 = (\Sigma\,(X_{1j} - M_1)^2 + \Sigma\,(X_{2j} - M_2)^2) / (n-1)$ be the est. var.
  - Let m and s be the observed value of M and S, respectively
  - Assumptions: M normal dist (reasonable) and $s = \sigma$ (naïve!)
  - Normalization: Define $Z = \sqrt{n} \cdot M / (2s)$ … then $\mathbf{E}\, Z = 0$ and $\mathbf{var}\, Z = 1$, and hence $Z \sim N(0, 1)$
  - P-value: $Pr(Z \geq x)$ where $x = \sqrt{n} \cdot m / (2s)$

    The probability that Z is ≥ it's observed value

Z-Test and T-Test 11a/12

- The T-Test improved slide, for use in future course
  - Given two series X_1 and X_2 of a total of n measurements

    Common unknown mean $\mu = \mathbf{E}\, X_{ij}$ and variance $\sigma = \mathbf{var}\, X_{ij}$
  - Let $M = M_1 - M_2$ be the difference of the means of X_1 and X_2
  - Let $S^2 = (\Sigma\, (X_{1j} - M_1)^2 + \Sigma\, (X_{2j} - M_2)^2) / (n-1)$ be the est. var.
  - Let m and s be the observed value of M and S, respectively
  - Assumptions: M normal dist and S^2 has χ^2 dist (both reasonable)
  - Normalization: Define $Z = \sqrt{n} \cdot M / (2\sigma) \sim N(0,1)$ and $V = S^2/\sigma^2 \cdot n \sim \chi^2(n)$ … then $T = \sqrt{n} \cdot M / (2S) = Z / \sqrt{(V/n)} \sim$ t-dist(n)
  - P-value: $\Pr(T \geq x)$ where $x = \sqrt{n} \cdot m / (2s)$

    The probability that T is ≥ it's observed value

Z-Test and T-Test 12/12

two-sided: $Pr(Z \le -x \text{ OR } Z \ge x)$

Both tests say: NOT statistically significant !

$n = 8$

- Back to our rolling dice example
  - Recall our two series of dice rolls

    A : 1 , 3 , 3 , 5 $\quad \mu_1 = \frac{1+3+3+5}{4} = 3, \quad \sigma_1^2 = \frac{(1-3)^2+(3-3)^2+(3-3)^2+(5-3)^2}{4} = \frac{4+4}{4} = 2.0$

    B : 6 , 6 , 4 , 4 $\quad \mu_2 = \frac{6+6+4+4}{4} = 5, \quad \sigma_2^2 = \frac{1^2+1^2+1^2+1^2}{4} = 1.0$

  - Observed difference of means $\Delta\mu$ is : $|3-5| = 2$
  - Observed estimated variance σ^2 is : $\sigma_1^2 + \sigma_2^2 = 3, \quad \sigma = \sqrt{3}$
  - Value x of $\sqrt{n} \cdot \Delta\mu / (2\sigma)$ is : $\sqrt{8} \cdot 2 / 2\sqrt{3} = \sqrt{8/3} \approx 1.63$
  - **Z-test:** p-value $Pr(Z \ge x)$ is : $\approx 5.2\%$ $\quad$ two-sided $\approx 10.4\%$
  - **T-test:** p-value $Pr(T \ge x)$ is : $\approx 7.1\%$ $\quad \approx 14.2\%$

    For "two-sided" test, simply multiply p-value by 2

This is a mistake: σ^2 should be the average of $\sigma_1{}^2$ and $\sigma_2{}^2$, not the sum
hence σ^2 = 1.5 and not 3 … the numbers below change accordingly

References

- Further reading

  Smucker, Allan, Carterette: A Comparison of Statistical Significance Tests for IR Evaluation, CIKM 2007

  http://ciir-publications.cs.umass.edu/getpdf.php?id=744

- Wikipedia
  - http://en.wikipedia.org/wiki/Statistical_hypothesis_testing
  - http://en.wikipedia.org/wiki/P-Value
  - http://en.wikipedia.org/wiki/Z-test
  - http://en.wikipedia.org/wiki/Student's_t-test
  - http://en.wikipedia.org/wiki/Student's_t-distribution