ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

Μηχανική μάθηση

Ενότητα 3: Hypothesis testing basics

Ιωάννης Τσαμαρδίνος
Τμήμα Επιστήμης Υπολογιστών

Hypotheses in Machine Learning

- Regarding the application of a model
  - Is the model really predictive?
  - Is the performance of the model above a certain threshold?
- Regarding model and algorithm comparison
  - Is model *A* more predictive than model *B*?
  - Is algorithm *A* more predictive than algorithm *B*?
- Within algorithms?
  - Is variable *A* independent of the target *Y*?
  - Does the Naïve Bayes assumption holds for variables *X*,*Z* and class Y?
- Etc.

Stochasticity of Hypotheses

- May always hold, but there is uncertainty about it unless we check all cases
  - Algorithm *A* is always better than algorithm *B* in all problems we have seen, but what are the chances this is true for all problems?
  - Naïve Bayes assumption holds for variables *X*,*Z* and class Y but we have not checked of the data population
- May only hold probabilistically
  - Algorithm *A* is most often better than algorithm *B*

Hypotheses in Science

- Mathematics:
  - Is P=NP?
- Natural Sciences and Other Sciences
  - Pharmacology: Does medicine A help patients recover more quickly from disease B?
  - Biology: Does gene A "cause" disease B?
  - Computer Science: Is algorithm A more time-efficient than algorithm B?

Proving Your Hypothesis

- By Reaching a contradiction in mathematics
  - We already know a set of axioms and theorems, say K
  - We want to show the theorem (hypothesis) H, e.g., "H: P=NP"
  - We show K, $\neg$H $\Rightarrow$ False (contradiction)
  - Thus, if we trust that K holds indeed
    - $\neg$H cannot hold
    - H must hold

"Proving Your Hypothesis"

- Mathematics
  1. We already know a set of axioms and theorems, say K
  2. We want to show the theorem (hypothesis) H, e.g., "H: P=NP"
  3. We show K, ¬H ⇒ False (contradiction)
  4. Thus, if we trust that K holds indeed, ¬H cannot hold and H must hold

- Physical World or Stochastic Situations in General
  1. We already "know" K
  2. We want to show a hypothesis H, e.g., "H: medicine A reduces the mortality of disease B"
  3. We gather data from the real world. We show that K, ¬H makes it very unlikely to observe our data
  4. We conclude that ¬H is very unlikely. We reject ¬H, we accept H

Example

- Ternary predictors $\{X_1, X_2\}$, binary target output Y
- Should we use the predictors or not?
  - We decide to use them only if they are dependent with Y
  - Is this always correct?
- Form the Hypothesis
  - H: X_1 is dependent of Y
- Data gathering:
  - Observe sample $\{\langle a,b,y \rangle_i, a \in Domain(X_1), b \in Domain(X_2), y \in Domain(Y)\}$

Example (hypothetical) Data

Example	X_1	X_2	Y
1 (x_1)	0	1	1
2	0	1	1
3	1	2	2
4	1	0	1
5	0	0	1
6	2	0	2
7	1	1	1
8	2	1	2
9	2	2	2
10	0	0	1
11	1	1	1
12	2	2	2
13	2	0	2
14	0	1	1
15 (x_{15})	1	2	2

Do We Accept Our Hypothesis?

- Seems they are dependent (why?)
- Well, maybe. But, maybe we are just unlucky. Maybe $\neg H$ does hold and still got these numbers.
- Statistics to the rescue:
  - If $\neg$H then we have a 0.0009 chance to observe these means in our data
- We conclude to accept *H*

X_1 \ Y	1	2
0	5	0
1	3	2
2	0	5

The Null Hypothesis

- The hypothesis we hope to accept is called the *Alternative Hypothesis*
  - *Sometimes denoted as H_1*

- The hypothesis we hope to reject, the negation of the Alternative Hypothesis, is called the *Null Hypothesis*
  - Usually denoted by H_o

A Bit of Notation

- Random variables are denoted with a capital letter, e.g., X
- Observed quantities of random variables are denoted with their corresponding small letter x

The Need for a Test Statistic

- Let $x_1, \ldots, x_n$ be the vectors with our data (e.g., each patient vector)
  - $x = \langle x_1, \ldots, x_n \rangle$
  - x is the matrix with our data
- Ideally, we would like to calculate:
  - P(obtaining data similar to $x \mid H_0$)
  - (Why not calculate $P(x \mid H_0)$?)
- This is typically very difficult

Test Statistics

- A test statistic is a function of our data: $T(X)$,
  - *e.g., if* X contains a single quantity (variable) $T(X)$ the mean value of X
- T is a random variable (since it depends on our random variables, the data)
- We denote with $t_o = T(x)$ is the observed value of T in our data
- Instead of calculating
  - P(obtaining data similar to $X \mid H_0$)
- we calculate
  - $P(T$ similar to $t_o \mid H_0)$
- If $P(T$ similar to $t_o \mid H_0)$ is very low, while $P(T$ similar to $t_o \mid H_1)$ is high, we reject H_0

Standard Single Hypothesis Testing

1. Form the Null and Alternative Hypothesis
2. Gather related data
3. Find a suitable test statistic T
4. Find the distribution of T
5. Depending on the distribution of T and the observed $t_o = T(x)$
   - decide to reject or not H_0

1. Form the Null Hypothesis

- H_0 : X_1 is independent of Y
- H_1 : X_1 is dependent of Y
- Mathematically?
  - $p_{ij} = P(X_1 = i, Y=j)$
  - $p_{i+} = P(X_1 = i)$
  - $p_{+j} = P(Y = j)$
- H_0 : $\forall i,j,\ p_{ij} = p_{i+} \cdot p_{+j}$,
- H_1 : $\exists i,j$ s.t., $p_{ij} \neq p_{i+} \cdot p_{+j}$

2. Gather Related Data

- If you don't have data …
  - Well, then become a philosopher

3. Find a suitable test statistic *T*

- Let's have another look at the data

Counts

X_1 \ Y	1	2	Out of
0	5	0	5
1	3	2	5
2	0	5	5
	8	7	15

Estimated Probs

X_1 \ Y	1	2
0	5/15	0/15
1	3/15	2/15
2	0/15	5/15

3. Find a suitable test statistic *T*

Counts

Actual Data

X_1 \ Y	1	2	Out of
0	5	0	5
1	3	2	5
2	0	5	5
	8	7	15

Estimated Probs

X_1 \ Y	1	2	
0	5/15	0/15	5/15
1	3/15	2/15	5/15
2	0/15	5/15	5/15
	8/15	7/15	

Expected under H_0?

X_1 \ Y	1	2	Out of
0			
1			
2			

?

3. Find a suitable test statistic *T*

Actual Data

Counts

X_1 \ Y	1	2	Out of
0	5	0	5
1	3	2	5
2	0	5	5
	8	7	15

Estimated Probs

X_1 \ Y	1	2	
0	5/15	0/15	5/15
1	3/15	2/15	5/15
2	0/15	5/15	5/15
	8/15	7/15	

Expected under H_0?

X_1 \ Y	1	2	Out of
0			
1			
2			

$$\frac{7}{15}\cdot\frac{5}{15}\cdot 15$$

3. Find a suitable test statistic *T*

- Observed counts of cases falling in cell *i,j*: O_{ij}
- Observed counts of cases falling in row *i*: O_{i+}
- Observed counts of cases falling in column *j*: O_{+j}
- Expected counts of cases falling in cell *i,j*: E_{ij}
- Under the Null Hypothesis: $E_{ij} = n(O_{i+} / n)(O_{+j} / n)$

3. Find a suitable test statistic *T*

Actual Data

X_1 \ Y	1	2	Out of
0	$5=O_{01}$	$0=O_{02}$	$5=O_{0+}$
1	$3=O_{11}$	$2=O_{12}$	$5=O_{1+}$
2	$0=O_{21}$	$5=O_{22}$	$5=O_{2+}$
	$8=O_{+1}$	$7=O_{+2}$	$15=n$

Expected under H_0

X_1 \ Y	1	2	Out of
0	$O_{0+}O_{+1}/n$ 40/15	$O_{0+}O_{+1}/n$ 35/15	$5=O_{0+}$
1	$nO_{0+}O_{+1}/n$ 40/15	$O_{0+}O_{+1}/n$ 35/15	$5=O_{1+}$
2	$O_{0+}O_{+1}/n$ 40/15	$O_{0+}O_{+1}/n$ 35/15	$5=O_{2+}$
	$8=O_{+1}$	$7=O_{+2}$	$15=n$

3. Find a suitable test statistic *T*

- Is there a way to summarize how close our observed data match what we expected assuming H_0 ?
  - The larger the difference O_{ij} to E_{ij} the less likely H_0 seems
  - Any ideas?

3. Find a suitable test statistic *T*

- Is there a way to summarize how close our observed data match what we expected assuming H_0 ?
  - The larger the difference O_{ij} to E_{ij} the less likely H_0 seems
  - Any ideas? Yes two!

$$T(X) = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (called\ X^2)$$

$$T(X) = 2\sum_{i,j} O_{ij} \log \frac{O_{ij}}{E_{ij}} \quad (called\ G^2)$$

4. Find the distribution of *T*

- Go to your statistician friend and ask him, open a statistics book, prove it yourself (!) etc.
- **if**
  a) Each example was sampled independently
  b) and **THE NULL HYPOTHESIS HOLDS**
- **then**

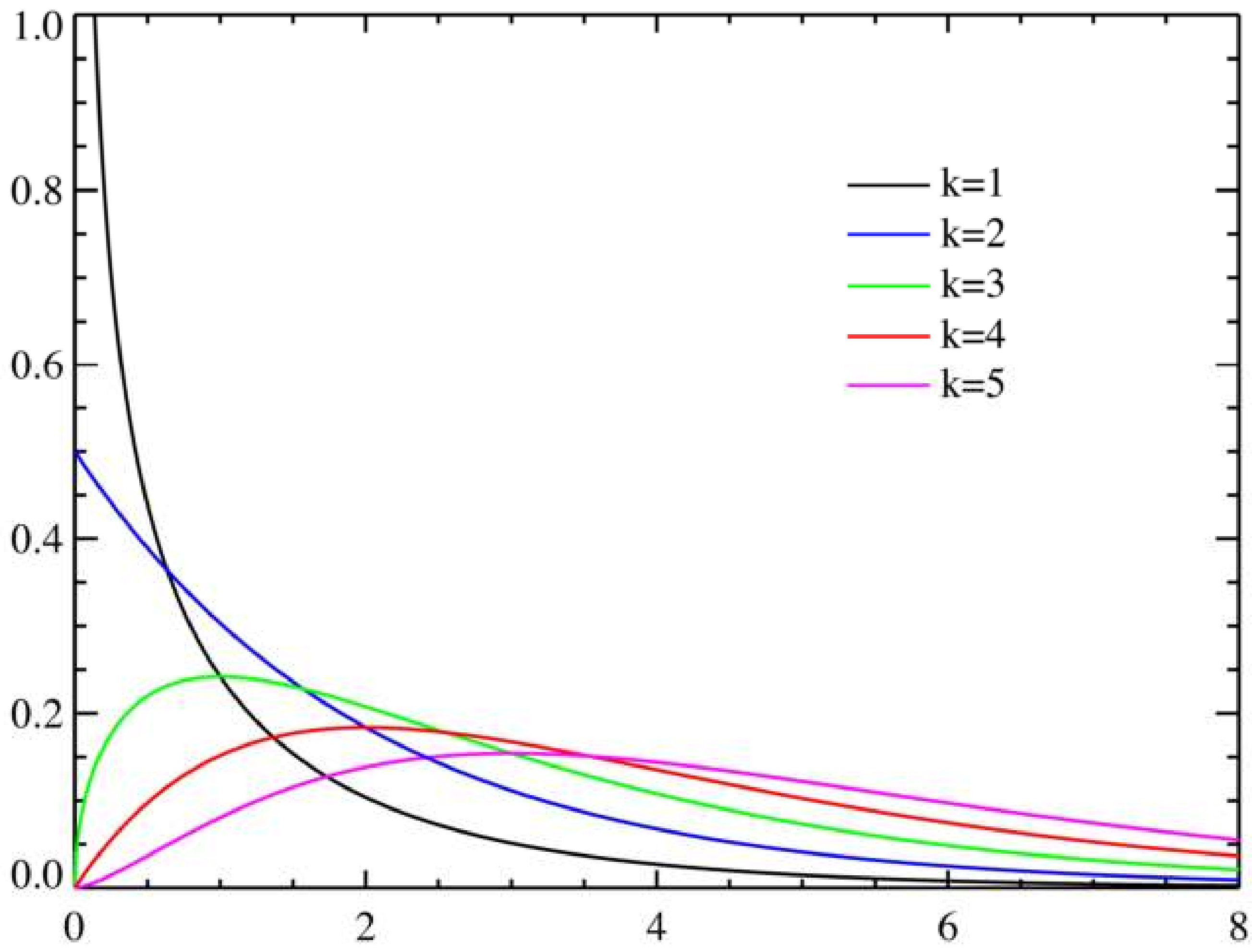
4. Find the distribution of *T*

- $T(X)=t$ (both for G^2 *and for* X^2) has a probability density function of:

$$p(t \mid H_0) = \frac{x^{(df-2)/2} e^{-x/2}}{2^{v/2} \Gamma(df/2)}$$

- where *df* is called the degrees of freedom of the test
  - the number of parameters that are free to vary
  - in our case it is (I-1)·(J-1) = (3-1) · (2-1) = 2
    - I = the number of possible values of X_1
    - J = the number of possible values of *Y*

The χ^2 distribution



5. Decide to Reject Or NOT

- Depending on the distribution of T and the observed $t_o = T(x)$
  - decide to reject or not H_0
- Should we or should we not
  - In our example the obsrved G^2 , $T(x) = 13.9976$
- If yes, what's the probability we err?

Decide on a Rejection Region

- Decide on a rejection region Γ in the range of our statistic
- If $t_o \in \Gamma$, then reject H_0
- If $t_o \notin \Gamma$, then do not reject H_0
  - accept H_1 ?
- Since we know the pdf of T ***when the null hypothesis holds*** *we can calculate* $P(T \in \Gamma \mid H_0)$

Decide on a Rejection Region

- If P($T \in \Gamma \mid H_0$) is too low, we know we are safely rejecting H_0
- However, we would also like P($T \in \Gamma \mid H_1$) to be high
- What should be our rejection region in our example?

Decide on a Rejection Region

- where extreme values of t_o are:
  - unlikely to come from when H_0 is true
  - could come with high probability, when H_0 is false
- $P(T \in \Gamma \mid H_0)$ is the area of the shaded region (can be calculated)

Typical Rejection Regions

- Typically, test statistics get extreme (too high or too low) values with low probability, when H_0 holds

Rejection Procedure

- We pre-select a probability threshold α
- We find a rejection region $\Gamma = \{t: |t| > c\}$ such that
  - $P(T \in \Gamma \mid H_0) = \alpha$
- We decide
  - Reject H_0 if $t_o \in \Gamma$ (recall: t_o is the observed T in our data)
  - *Accept* H_0 otherwise
- What values do we usually use for α in science?
  - 0.05 is the most typical
  - smaller ones are also used: 0.01 , 0.001 etc.
- When $t_o \in \Gamma$ we say the finding is statistically significant at significance level α
- In our example for $\alpha = 0.05$, $df=2$, the rejection region is any observed G^2 greater than 5.9915

How to Construct a Test

- Find a test statistic *T* that easily takes extreme values when H_0 is not true
- Find how to theoretically calculate the distribution of *T*
  - hopefully, without too many other assumptions

- Our example described the application of what is called a χ^2 test of independence

Statistical Errors

- Assuming all the other assumptions of the test hold what types of errors can we make?

Decision / Truth	Accept H_0	Reject H_0
H_0		Type I error
H_1	Type II error	

Decision / Truth	Accept H_0	Reject H_0
H_0	True Positive	False Positive
H_1	False Negative	True Negative

Statistical Errors

Decision / Truth	Accept H_0	Reject H_0
H_0	$P(T \notin \Gamma, H_0)$	$P(T \in \Gamma, H_0)$
H_1	$P(T \notin \Gamma, H_1)$	$P(T \in \Gamma, H_1)$

- Out of these, which ones can we compute?

Statistical Errors

Decision Truth	Accept H_0	Reject H_0
H_0	$P(t \notin \Gamma \mid H_0)\ P(H_0)$	$P(t \in \Gamma \mid H_0)\ P(H_0)$
H_1	$P(t \notin \Gamma \mid H_1)\ P(H_1)$	$P(t \in \Gamma \mid H_1)\ P(H_1)$

- We showed we can compute $P(T \in \Gamma \mid H_0)$
- Under additional assumptions, we can calculate $P(T \notin \Gamma \mid H_1)$
- We could compute the rest under more assumptions
- Requires a different semantic interpretation of probability, called Bayesian Statistics

Statistical Errors

- Type I error $\leq$ $P(T \in \Gamma \mid H_0)$
- Type II error $\leq P(T \notin \Gamma \mid H_1)$
- $P(T \in \Gamma \mid H_1) = 1 - P(T \notin \Gamma \mid H_1)$ is called the statistical power of the test:
  - probability of accepting H_1 given that H_1 is true

Minimize Statistical Error

- Problem 1: minimize the total number of errors
- Problem 2: minimize a weighted sum of errors
  - desirable if one type of error is more serious than another
- Problem 3 (standard statistics):
  - minimize Type II errors
  - while Type I errors are below a threshold

Minimize Statistical Errors

- Find test statistics and rejection regions s.t.
- Guarantee Type I errors are below a given level α
  - Type I error $\leq P(T \in \Gamma \mid H_0) = \alpha$
  - we say the test controls Type I error rate
- While minimizing Type II errors
  - that is maximizing power : $P(T \in \Gamma \mid H_1)$

Statistical Significance

- We set a threshold α (=0.05 often)
- We find a value c, s.t. $P(|T| \geq c \mid H_0) = \alpha$
  - this defines our rejection region
- We reject H_0 if $|t_o| \geq c$, thus we control the Type I error rate to be less than α
- α is called the significance level of the test procedure

p-values

- For a significance level *a*, we report whether we reject H_0 or not
- Or better,
- We calculate the minimum *a* that would still have rejected H_0 (p-value)

- $p = P(|T| \geq |t_0| \;\; | H_0)$

- In the example,
  - we can reject the hypothesis of indepencene at level 0.05
  - the *p-value* is 0.0009

p-value

- The probability of obtaining a test statistic as extreme ***or more*** as the one we have obtained, ***given*** H_0

Regarding the Choice of a Test

- When we cannot reject H_0 does not mean H_1 holds
- It could be that we don't have enough power, i.e.,
  - H_1 is not that "different enough" from H_0 to distinguish it with the given sample size we have
- At of all possible tests for a hypothesis choose the one with the maximum power
- tests with more assumptions are typically more powerful
  - and less likely to be applicable

Disclaimer

- We have simplified certain aspects of hypothesis testing
- What we said holds for what is called simple tests

Τέλος Ενότητας

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Κρήτης**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σημειώματα

Σημείωμα αδειοδότησης (1)

Σημείωμα αδειοδότησης (2)

- Ως **Μη Εμπορική** ορίζεται η χρήση:
  - που δεν περιλαμβάνει άμεσο ή έμμεσο οικονομικό όφελος από την χρήση του έργου, για το διανομέα του έργου και αδειοδόχο
  - που δεν περιλαμβάνει οικονομική συναλλαγή ως προϋπόθεση για τη χρήση ή πρόσβαση στο έργο
  - που δεν προσπορίζει στο διανομέα του έργου και αδειοδόχο έμμεσο οικονομικό όφελος (π.χ. διαφημίσεις) από την προβολή του έργου σε διαδικτυακό τόπο

- Ο δικαιούχος μπορεί να παρέχει στον αδειοδόχο ξεχωριστή άδεια να χρησιμοποιεί το έργο για εμπορική χρήση, εφόσον αυτό του ζητηθεί.

Σημείωμα Αναφοράς

Copyright Πανεπιστήμιο Κρήτης, Ιωάννης Τσαμαρδίνος 2015. «Μηχανική Μάθηση. Hypothesis testing basics». Έκδοση: 1.0. Ηράκλειο 2015. Διαθέσιμο από τη δικτυακή διεύθυνση: https://opencourses.uoc.gr/courses/course/view.php?id=362.

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.