# Ανάκτηση Πληροφορίας

**Ενότητα 5**: Μοντελοποίηση: Πιθανοκρατικό Μοντέλο

Απόστολος Παπαδόπουλος
Τμήμα Πληροφορικής



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

# Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

# Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «Ανοικτά Ακαδημαϊκά Μαθήματα στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης» έχει χρηματοδοτήσει μόνο την 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.

# Το Πιθανοκρατικό Μοντέλο



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

# Περιεχόμενα ενότητας

1. Το Πιθανοκρατικό Μοντέλο



# Το Πιθανοκρατικό Μοντέλο

# Κλασικά Μοντέλα Ανάκτησης

Τρία είναι τα, λεγόμενα, κλασικά μοντέλα ανάκτησης:

**Λογικό (Boolean)** που βασίζεται στη Θεωρία Συνόλων

**Διανυσματικό (Vector)** που βασίζεται στη Γραμμική Άλγεβρα

**Πιθανοκρατικό (Probabilistic)** που βασίζεται στη Θεωρία Πιθανοτήτων

Τα δύο πρώτα μοντέλα έχουν ήδη εξεταστεί. Τονίζεται επίσης, ότι το Διανυσματικό και το Πιθανοκρατικό έχουν σημαντική επικάλυψη αν και στηρίζονται σε εντελώς διαφορετικές θεωρίες.

# Πιθανοκρατικό Μοντέλο-1

- Στόχος: να ορίσουμε το IR πρόβλημα σε πιθανοτικό πλαίσιο.
- Για κάθε user query υπάρχει ένα ιδανικό σύνολο κειμένων που το ικανοποιεί.
- Η ερώτηση επεξεργάζεται με βάση τις ιδιότητες αυτού του συνόλου.
- Ποιες είναι όμως αυτές οι ιδιότητες;
- Αρχικά γίνεται μία πρόβλεψη και στη συνέχεια η πρόβλεψη βελτιώνεται.

# Πιθανοκρατικό Μοντέλο-2

- Αρχικά επιστρέφεται ένα σύνολο εγγράφων.
- Ο χρήστης εξετάζει τα κείμενα αναζητώντας σχετικά κείμενα.
- Το σύστημα IR χρησιμοποιεί το feedback του χρήστη ώστε να προσδιοριστεί καλύτερα το ιδανικό σύνολο κειμένων.
- Η διαδικασία επαναλαμβάνεται.
- Η περιγραφή του ιδανικού συνόλου κειμένων πραγματοποιείται πιθανοτικά.

# Ανεξάρτητες Μεταβλητές και Πιθανότητα υπό Συνθήκη

Έστω *a,* και *b* δύο γεγονότα με πιθανότητες να συμβούν $P(a)$ και $P(b)$ αντίστοιχα.

**Ανεξάρτητα Γεγονότα**

Τα γεγονότα *a* και *b* είναι ανεξάρτητα αν και μόνο αν:

$$P(a \cap b) = P(b)\, P(a)$$

**Υπό Συνθήκη Πιθανότητα**

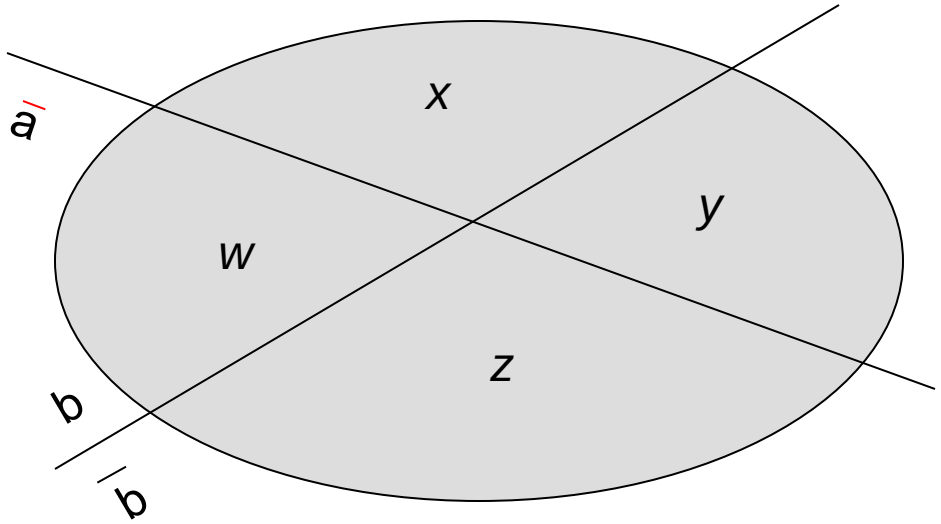
$P(a \mid b)$ είναι η πιθανότητα του *a* δεδομένου του *b*.

Τα γεγονότα $a_1, ..., a_n$ καλούνται υπό συνθήκη ανεξάρτητα αν και μόνο αν:

$$P(a_i \mid a_j) = P(a_i) \text{ για όλα τα } i \text{ και } j$$

# Παράδειγμα Ι

$\overline{a}$ είναι η άρνηση του γεγονότος $a$




$P(a) = x + y$

$P(b) = w + x$

$P(a \mid b) = x / (w + x)$

$\boldsymbol{P(a \mid b)\, P(b) = P(a \cap b) = P(b \mid a)\, P(a)}$

# Παράδειγμα ΙΙ

**Ανεξάρτητα γεγονότα**

Έστω *a* και *b* οι τιμές που φέρνουν δύο ίδια ζάρια. Ισχύει:

$$P(a=5 \mid b=3) = P(a=5) = {}^{1}/_{6}$$

**Μη ανεξάρτητα**

Έστω *a* και *b* οι τιμές που φέρνουν δύο ίδια ζάρια και *t* το άθροισμά τους. Τότε ισχύει:

$$t = a + b$$

$$P(t=8 \mid a=2) = {}^{1}/_{6}$$
$$P(t=8 \mid a=1) = 0$$

# Θεώρημα του Bayes

Έστω *a* και *b* δύο γεγονότα.

$P(a \mid b)$ είναι η πιθανότητα να συμβεί το γεγονός *a* δεδομένου ότι έχει συμβεί το γεγονός *b*.

**Θεώρημα Bayes**

$$P(a \mid b) = \frac{P(b \mid a)\, P(a)}{P(b)}$$

**Ισχύει επίσης ότι:**

$$P(a \mid b)\, P(b) = P(a \cap b) = P(b \mid a)\, P(a)$$

# Θεώρημα Bayes: παράδειγμα

Example

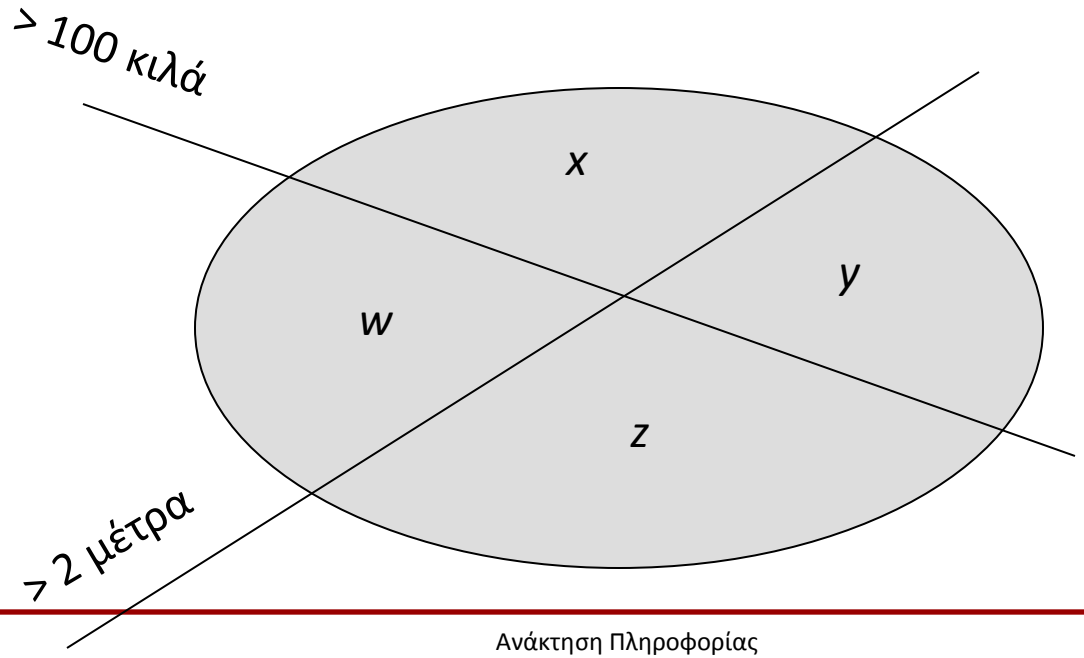
*a* βάρος πάνω από 100 κιλά

*b* ύψος πάνω από 2 μέτρα.

$P(a \mid b) = x / (w+x) = x / P(b)$

$P(b \mid a) = x / (x+y) = x / P(a)$

$x = P(a \cap b)$



# Αρχή Πιθανοκρατικής Βαθμολόγησης

"If a reference retrieval system's response to each request is a ranking of the documents in the collections in order of decreasing **probability of usefulness** to the user who submitted the request, where the probabilities are estimated as accurately a possible on the basis of whatever data is made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data."

Εάν η απάντηση ενός συστήματος ανάκτησης σε κάθε ερώτημα είναι μία λίστα εγγράφων ταξινομημένη με φθίνουσα διάταξη ως προς την **πιθανότητα σχετικότητας** του κάθε εγγράφου ως προς το χρήστη, όπου οι πιθανότητες υπολογίζονται όσο γίνεται ακριβέστερα με βάση τα δεδομένα που είναι διαθέσιμα, η συνολική αποτελεσματικότητα του συστήματος θα είναι η καλύτερη δυνατή.

*W.S. Cooper*

# Πιθανοκρατική Βαθμολόγηση

“Για ένα δεδομένο ερώτημα, εάν γνωρίζουμε κάποια από τα σχετικά έγγραφα, οι όροι που εμφανίζονται σε αυτά θα πρέπει να έχουν μεγαλύτερη βαρύτητα κατά την αναζήτηση άλλων σχετικών εγγράφων. Κάνοντας διάφορες παραδοχές σχετικά με την κατανομή των όρων και χρησιμοποιώντας το θεώρημα του Bayes είναι δυνατόν να υπολογίσουμε τα βάρη αυτά.”

*Van Rijsbergen*

# Βασικές Έννοιες-1

Η πιθανότητα ένα έγγραφο να είναι σχετικό ως προς το ερώτημα θεωρείται ότι εξαρτάται μόνο από τους όρους που περιέχονται στο έγγραφο και από τους όρους που περιέχονται στο ερώτημα.

Η σχετικότητα ενός εγγράφου *d* ως προς το ερώτημα *q* δεν εξαρτάται από τη σχετικότητα άλλων εγγράφων της συλλογής.

Για κάποιο ερώτημα *q* το σύνολο των σχετικών εγγράφων *R* είναι το ιδανικό σύνολο που μπορούμε να έχουμε ως απάντηση.

# Βασικές Έννοιες-2

Για ένα ερώτημα *q* και ένα έγγραφο *d* το πιθανοκρατικό μοντέλο χρειάζεται μία εκτίμηση για την πιθανότητα $P(R \mid d)$ που δηλώνει την πιθανότητα το έγγραφο *d* να είναι σχετικό ως προς το ερώτημα.

**Μέτρο Ομοιότητας:**

$S(q, d)$, ομοιότητα του εγγράφου *d* ως προς το ερώτημα *q*:

$$\frac{\text{πιθανότητα } d \text{ σχετικό}}{\text{πιθανότητα } d \text{ μη σχετικό}} = \frac{P(R \mid d)}{P(\neg R \mid d)}$$

Οι τιμές της $S(\ )$ μπορεί να είναι από πολύ μικρές έως πολύ μεγάλες και γι αυτό χρησιμοποιείται συνήθως ο λογάριθμος για την άμβλυνση των διαφορών.

# Βασικές Έννοιες-3

$$S(q, d) = \frac{P(R \mid d)}{P(\bar{R} \mid d)}$$

$$= \frac{P(d \mid R)\, P(R)}{P(d \mid \bar{R})\, P(\bar{R})} \quad \text{θεώρημα Bayes}$$

$$= \frac{P(d \mid R)}{P(d \mid \bar{R})} \times k \quad k \text{ μία σταθερά}$$

$P(d \mid R)$ είναι η πιθανότητα να διαλέξουμε τυχαία το $d$ από το $R$.

# Βασικές Έννοιες-4

**Ανάκτηση Δυαδικής Ανεξαρτησίας** (binary independence retrieval )

Τα βάρη των όρων είναι δυαδικά και οι όροι είναι ανεξάρτητοι μεταξύ τους (η παρουσία ή μη κάποιου όρου δεν επηρεάζει τους υπόλοιπους). Το βάρος ενός όρου σε ένα έγγραφο είναι είτε 1 (αν ο όρος περιέχεται στο έγγραφο) είτε 0 (σε διαφορετική περίπτωση).

Όπως και στο Λογικό αλλά και στο Διανυσματικό μοντέλο, η σχετικότητα ενός εγγράφου καθορίζεται από τους όρους που περιέχονται σε αυτό.

# Naïve Bayes

Έστω $\mathbf{x} = (x_1, x_2, \ldots x_n)$ το διάνυσμα του εγγράφου $d$ όπου
$x_i = 1$ αν ο i-οστός όρος περιέχεται στο έγγραφο, $x_i = 0$ διαφορετικά.

Η εκτίμηση της πιθανότητας $P(d \mid R)$ γίνεται χρησιμοποιώντας την πιθανότητα $P(\mathbf{x} \mid R)$

Εάν οι όροι είναι ανεξάρτητοι τότε:

$$P(\mathbf{x} \mid R) = P(x_1 \cap R)\, P(x_2 \cap R) \ldots P(x_n \cap R)$$

$$= P(x_1 \mid R)\, P(x_2 \mid R) \ldots P(x_n \mid R)$$

$$= \prod P(x_i \mid R)$$

$\boldsymbol{P(x_i \mid R)}$ είναι η πιθανότητα ο όρος $x_i$ να βρίσκεται σε ένα έγγραφο που επιλέγεται τυχαία από το σύνολο $R$.

Το μοντέλο αυτό είναι γνωστό και ως **Naive Bayes.**

# Συνάρτηση Ομοιότητας-1

$$S(q, d) = k \frac{\prod P(x_i \mid R)}{\prod P(x_i \mid \bar{R})}$$

Αφού το κάθε $x_i$ είναι 0 ή 1 έχουμε:

$$S = k \prod_{x_i = 1} \frac{P(x_{i=1} \mid R)}{P(x_{i=1} \mid \bar{R})} \prod_{x_i = 0} \frac{P(x_{i=0} \mid R)}{P(x_{i=0} \mid \bar{R})}$$

# Συνάρτηση Ομοιότητας-2

Για τους όρους που εμφανίζονται στο ερώτημα θέτουμε:

$$p_i = P(x_i = 1 \mid R)$$
$$r_i = P(x_i = 1 \mid \overline{R})$$

Για τους όρους που δεν εμφανίζονται στο ερώτημα έστω:

$$p_i = r_i$$

*όροι με* $q_i = 0$ *είναι ίσοι με* $p_i / r_i = 1$

$$S = k \prod_{x_i = q_i = 1} \frac{p_i}{r_i} \prod_{x_i = 0,\, q_i = 1} \frac{1 - p_i}{1 - r_i}$$

$$= k \prod_{x_i = q_i = 1} \frac{p_i (1 - r_i)}{r_i (1 - p_i)} \prod_{q_i = 1} \frac{1 - p_i}{1 - r_i}$$

*σταθερή ποσότητα για δεδομένο ερώτημα*

# Συνάρτηση Ομοιότητας-3

Με λογαρίθμηση της σχέσης και αγνοώντας σταθερούς παράγοντες η συνάρτηση ομοιότητας $S_{prob}(q,d)$ παίρνει τη μορφή:

$$S_{prob}(q,d) = \log\left(S(q,d)\right)$$

$$S_{prob}(q,d) = \sum_i \log \frac{p_i \cdot (1-r_i)}{r_i \cdot (1-p_i)}$$

Όπου η άθροιση αφορά στους όρους που βρίσκονται **και στο ερώτημα και στο έγγραφο**.

# Σχέση με το Διανυσματικό Μοντέλο

Στο Διανυσματικό μοντέλο ανάκτησης θεωρήστε ότι η *i*-οστή συνιστώσα του διανύσματος ενός εγγράφου ισούται με την ποσότητα

$$\log \frac{p_i \cdot (1 - r_i)}{r_i \cdot (1 - p_i)}$$

ενώ το διάνυσμα του ερωτήματος *q* ισούται με άσσους για τους όρους που ανήκουν στο ερώτημα και μηδενικά διαφορετικά.

Τότε, η συνάρτηση ομοιότητας $S_{prob}(q,d)$ ισούται με το εσωτερικό γινόμενο των δύο διανυσμάτων.

# Αρχική Εκτίμηση των P(xi | R)

Αρχικά θέτουμε τιμές στις πιθανότητες :

$$p_i = P(x_i \mid R) = c$$

$$r_i = P(x_i \mid R) = n_i / N$$

όπου:

$c$ είναι μία τυχαία σταθερά (π.χ., 0.5)
$n_i$ είναι το πλήθος των εγγράφων που περιέχουν τον i-οστό όρο
$N$ πλήθος εγγράφων συλλογής.

# Προσαρμογή Τιμών των P(xi | R)

Είναι προφανές ότι η αυθαίρετη ανάθεση τιμών δεν μπορεί να οδηγεί πάντα σε ικανοποιητικά αποτελέσματα. Για τη βελτίωση της ποιότητας των αποτελεσμάτων οι πρώτες εφαρμογές του Πιθανοκρατικού μοντέλου χρειαζόταν την παρέμβαση του χρήστη για την αναπροσαρμογή των τιμών.

**Εναλλακτικά** μπορεί να χρησιμοποιηθεί και αυτοματοποιη-μένος τρόπος. Αρχικά εκτελείται το ερώτημα με τις αρχικές εκτιμήσεις. Επιλέγονται τα $k$ καλύτερα έγγραφα. Έστω $k_i$ ο αριθμός των εγγράφων που περιέχουν τον i-οστό όρο. Θέτουμε:

$$p_i = P(x_i \mid R) = k_i / k$$

$$r_i = P(x_i \mid R) = (n_i - k_i) / (N - k)$$

# Πλεονεκτήματα-Μειονεκτήματα

**Πλεονεκτήματα**:

1. Απλό μοντέλο.
2. Τα κείμενα ταξινομούνται σε φθίνουσα διάταξη ως προς την πιθανότητα να είναι σχετικά.

**Μειονεκτήματα**:

1. Χρειάζεται να μαντέψουμε.
2. Δε λαμβάνεται υπ’ όψιν η συχνότητα εμφάνισης.
3. Θεωρεί ότι τα keywords είναι ανεξάρτητα.

# Βιβλιογραφία


Maron, M.E. and Kuhns, J.L. “On Relevance, Probabilistic Indexing and Information Retrieval”. *Journal of the ACM*, 7, pp. 216-244, 1960.

Robertson, S.E. “The Probability Ranking Principle in IR”, *Journal of Documentation*, 33, pp. 294-304, 1977.

Fuhr, N. “Probabilistic Models in Information Retrieval”. *Computer Journal*. 35 (3), pp. 243-55, 1992.

Baeza-Yates, R. and Ribeiro-Neto, B. “*Modern Information Retrieval”*, Addison Wesley, 1999.

Manning, C.D. and Raghavan, P. and Schutze, H. “*An Introduction to Information Retrieval”*, Cambridge University Press, 2007.

# Σημείωμα Αναφοράς

# Σημείωμα Αδειοδότησης

Το παρόν υλικό διατίθεται με τους όρους της άδειας χρήσης Creative Commons Αναφορά - Μη Εμπορική Χρήση - Όχι Παράγωγα Έργα 4.0 [1] ή μεταγενέστερη, Διεθνής Έκδοση. Εξαιρούνται τα αυτοτελή έργα τρίτων π.χ. φωτογραφίες, διαγράμματα κ.λ.π., τα οποία εμπεριέχονται σε αυτό και τα οποία αναφέρονται μαζί με τους όρους χρήσης τους στο «Σημείωμα Χρήσης Έργων Τρίτων».


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

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

[1] http://creativecommons.org/licenses/by-nc-nd/4.0/

# Τέλος ενότητας

Επεξεργασία: <Μαυρίδης Απόστολος>
Θεσσαλονίκη, <Εαρινό εξάμηνο 2013-2014>



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

# 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.