ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Τεχνητή Νοημοσύνη

18η διάλεξη (2023-24)

Ίων Ανδρουτσόπουλος

http://www.aueb.gr/users/ion/

Οι διαφάνειες αυτής της διάλεξης βασίζονται:

- στο βιβλίο *Machine Learning* του T. Mitchell, McGraw-Hill, 1997,
- σε ύλη των διαλέξεων του μαθήματος Μηχανικής Μάθησης του A. Ng στο Πανεπιστήμιο Stanford (βλ. http://cs229.stanford.edu/).

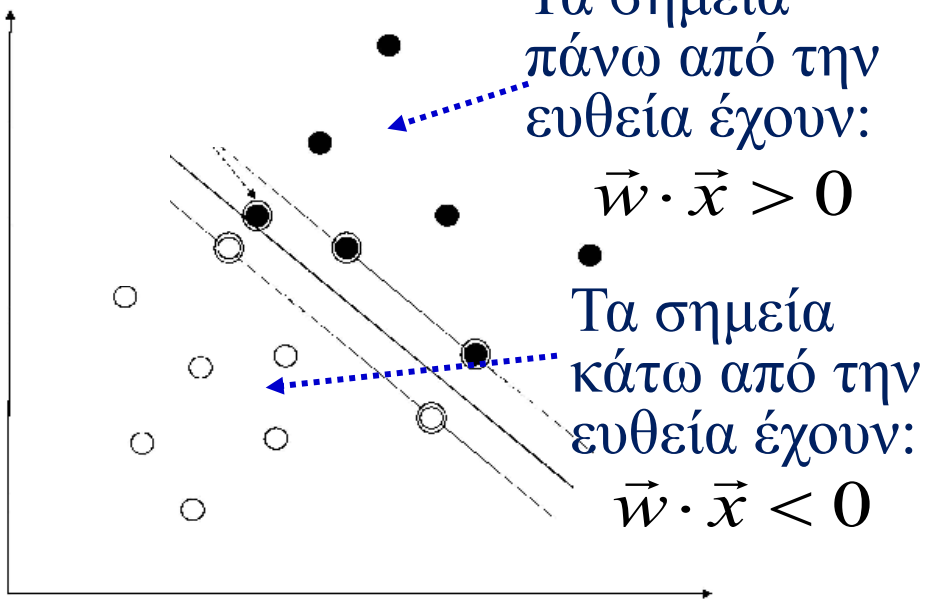
Τι θα ακούσετε σήμερα

- Γραμμικοί διαχωριστές.
- Ταξινομητές λογιστικής παλινδρόμησης.
- Μεγιστοποίηση πιθανοφάνειας με κατάβαση κλίσης.
- Ομαλοποίηση (regularization).
- Διαγνωστικοί έλεγχοι κατά τη χρήση επιβλεπόμενης μηχανικής μάθησης.
- Μέτρα αξιολόγησης για προβλήματα κατηγοριοποίησης και παλινδρόμησης.

Γραμμικοί διαχωριστές

- Για δύο ιδιότητες x_1, x_2, προσπαθούμε να μάθουμε **ευθεία** που **διαχωρίζει** τις δύο κατηγορίες.

$$w_2 x_2 + w_1 x_1 + w_0 = 0$$



- Γενικότερα, για ιδιότητες x_1, x_2, …, x_n προσπαθούμε να μάθουμε ένα **υπερ-επίπεδο** που να διαχωρίζει τις δύο κατηγορίες.

Θεωρούμε πάλι ότι $x_0 = 1$.

$$w_n x_n + \ldots + w_1 x_1 + w_0 = \sum_{l=0}^{n} w_l x_l = \vec{w} \cdot \vec{x} = 0$$

- **Απόφαση** κατάταξης:

$$C = sign(\vec{w} \cdot \vec{x})$$

Γραμμικοί διαχωριστές – συνέχεια

- Συχνά θέλουμε ο ταξινομητής να επιστρέφει και ένα **βαθμό βεβαιότητας**.
  - Π.χ. πόσο **πιθανό** θεωρεί να **ανήκει** ένα προς κατάταξη κείμενο με διάνυσμα $\vec{x}$ στη **μία** ή την **άλλη κατηγορία**.
- Η προσημασμένη **απόσταση** $d_{\vec{w}}(\vec{x})$ από το υπερ-επίπεδο διαχωρισμού **δεν** είναι **καλός** βαθμός βεβαιότητας.

$$d_{\vec{w}}(\vec{x}) = \frac{\vec{w} \cdot \vec{x}}{\|\vec{w}\|}$$

Χωρίς το w_0.

  - **Δεν** είναι **περιορισμένη** στο **[0, 1].**
  - Για **μεγάλες** (θετικές ή αρνητικές) **αποστάσεις** θέλουμε η βεβαιότητα να **τείνει στο 1**.
  - Για **μικρές αποστάσεις** θέλουμε η βεβαιότητα να **τείνει στο 0**.

Σιγμοειδής συνάρτηση (logistic function)

- Στην περίπτωσή μας το t θα είναι η **προσημασμένη** (και μη κανονικοποιημένη) **απόσταση** από το **υπερ-επίπεδο** διαχωρισμού:

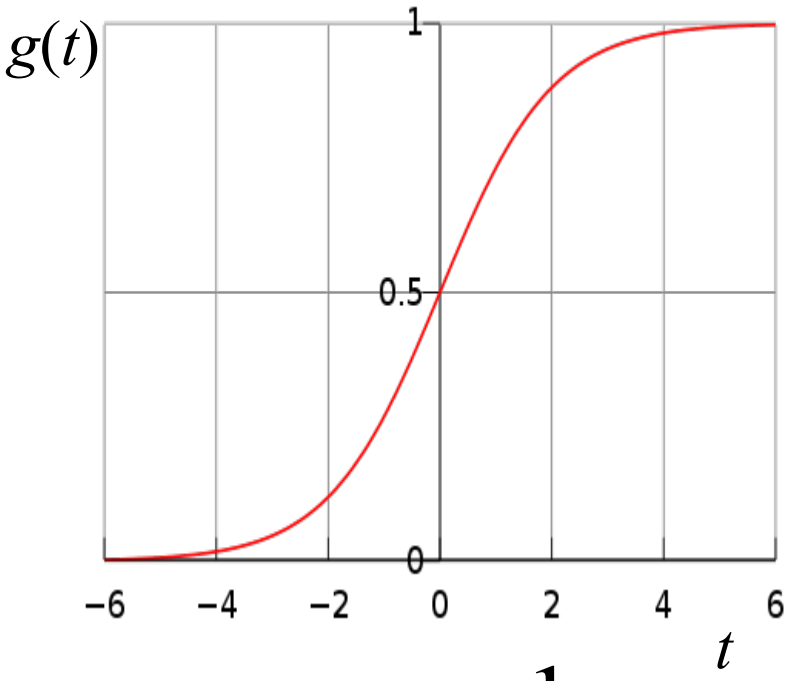
$$t = \vec{w} \cdot \vec{x}$$

- **Πιθανότητα** το $\vec{x}$ να ανήκει στη **θετική** κατηγορία:

$$P(c_+ \mid \vec{x}) = \frac{1}{1 + e^{-\vec{w} \cdot \vec{x}}}$$

- **Πιθανότητα** να ανήκει στην **αρνητική** κατηγορία:

$$P(c_- \mid \vec{x}) = 1 - P(c_+ \mid \vec{x})$$



Πηγή εικόνας: http://en.wikipedia.org/wiki/Logistic_function

$$g(t) = \frac{1}{1 + e^{-t}}$$

Ταξινομητές λογιστικής παλινδρόμησης (logistic regression classifiers)

$$P(c_+ \mid \vec{x}) = \frac{1}{1+e^{-\vec{w}\cdot\vec{x}}},\ P(c_- \mid \vec{x}) = 1 - P(c_+ \mid \vec{x}) = \frac{e^{-\vec{w}\cdot\vec{x}}}{1+e^{-\vec{w}\cdot\vec{x}}}$$

- Κατά την **εκπαίδευση**, επιλέγουν το $\vec{w}$ που κάνει τον ταξινομητή πιο βέβαιο ότι τα **παραδείγματα εκπαίδευσης** ανήκουν στις **σωστές κατηγορίες**.
  - Μεγιστοποιούν τη (δεσμευμένη) «**πιθανοφάνεια**» των παραδειγμάτων.

$$L(\vec{w}) = P(y^{(1)},\ldots,y^{(m)} \mid \vec{x}^{(1)},\ldots,\vec{x}^{(m)};\vec{w})$$

Οι σωστές κατηγορίες των παραδειγμάτων εκπαίδευσης.

Τα παραδείγματα εκπαίδευσης.

Μεγιστοποίηση πιθανοφάνειας

- Θεωρώντας ότι τα παραδείγματα εκπαίδευσης έχουν επιλεγεί από τον **ίδιο πληθυσμό** και είναι **ανεξάρτητα**:

$$L(\vec{w}) = P(y^{(1)},\ldots,y^{(m)} \mid \vec{x}^{(1)},\ldots,\vec{x}^{(m)};\vec{w})$$

$$= \prod_{i=1}^{m} P(y^{(i)} \mid \vec{x}^{(i)};\vec{w})$$

- Αντί να μεγιστοποιήσουμε την $L(\vec{w})$, βολεύει να μεγιστοποιήσουμε τη (γνησίως αύξουσα):

$$l(\vec{w}) = \log L(\vec{w}) = \sum_{i=1}^{m} \log P(y^{(i)} \mid \vec{x}^{(i)};\vec{w})$$

  - Είναι **κοίλη συνάρτηση**, δεν υπάρχει κίνδυνος να φτάσω σε τοπικό μέγιστο.

Μεγιστοποίηση πιθανοφάνειας – συνέχεια

- Αν παραστήσουμε τις (σωστές) κατηγορίες με $y = 1$ (**θετική** κατηγορία) και $y = 0$ (**αρνητική**), τότε:

$$P(y \mid \vec{x}; \vec{w}) = P(c_+ \mid \vec{x}; \vec{w})^y \cdot P(c_- \mid \vec{x}; \vec{w})^{(1-y)}$$

  - Για $y = 1$ (θετική κατηγορία), ο 2ος όρος εξαφανίζεται.
  - Για $y = 0$ (αρνητική), ο 1ος όρος εξαφανίζεται.

- Οπότε:

$$l(\vec{w}) = \sum_{i=1}^{m} \log P(c_+ \mid \vec{x}^{(i)}; \vec{w})^{y^{(i)}} + \log P(c_- \mid \vec{x}^{(i)}; \vec{w})^{(1-y^{(i)})}$$

$$= \sum_{i=1}^{m} y^{(i)} \log P(c_+ \mid \vec{x}^{(i)}; \vec{w}) + (1 - y^{(i)}) \log P(c_- \mid \vec{x}^{(i)}; \vec{w})$$

Μεγιστοποίηση πιθανοφάνειας – συνέχεια

Τώρα **μεγιστοποιούμε** το $l(\vec{w})$,
αντί να ελαχιστοποιούμε το $E(\vec{w})$.

- Με **<u>ανάβαση</u> κλίσης**:

$$\vec{w} \leftarrow \vec{w} + \eta \cdot \nabla l(\vec{w})$$

καταλήγουμε στον κανόνα ενημέρωσης:

$$w_l \leftarrow w_l + \eta \cdot \sum_{i=1}^{m} [y^{(i)} - P(c_+ \mid \vec{x}^{(i)})] \cdot x_l{}^{(i)}$$

- Εναλλακτικά μπορούμε να χρησιμοποιήσουμε π.χ. **στοχαστική ανάβαση κλίσης**.
  - Δεν υπάρχει κλειστή λύση.

Ομαλοποίηση (regularization)

- Στην πράξη αντί για το:

$$l(\vec{w}) = \sum_{i=1}^{m} \log P(y^{(i)} \mid \vec{x}^{(i)}; \vec{w})$$

συνήθως μεγιστοποιούμε το:

$$l(\vec{w}) - \lambda \cdot \|\vec{w}\|^2 = l(\vec{w}) - \lambda \cdot \sum_{l=0}^{n} w_l^2$$

L2 regularization ("ridge regression")

δηλ. επιβραβεύουμε υποψήφια $\vec{w}$ με πολλά μικρά βάρη.

Το L1 regularization ("lasso regression") χρησιμοποιεί τη νόρμα L1, δηλ. προσθέτει $-\lambda \sum_{l=0}^{n} |w_l|$. Οδηγεί σε πιο αραιά $\vec{w}$ (με πολλά μηδενικά).

- Υπάρχει έτσι **μικρότερος κίνδυνος υπερ-εφαρμογής**.
  - Π.χ. αν πολλά βάρη w_l είναι πολύ μικρά, οι αντίστοιχες ιδιότητες ουσιαστικά δεν χρησιμοποιούνται. Με **λιγότερες ιδιότητες** έχουμε **μικρότερο κίνδυνο υπερ-εφαρμογής**.
  - $\lambda > 0$. Η τιμή επιλέγεται με δοκιμές σε δεδομένα επικύρωσης.

Πολυωνυμική λογιστική παλινδρόμηση
(multinomial logistic regression)

- Επέκταση για **πολλές κατηγορίες** $c_1, c_2, \ldots, c_K$.
  - Ουσιαστικά μαθαίνουμε έναν **ξεχωριστό γραμμικό διαχωριστή για κάθε κατηγορία** c_i.

$$P(c_j \mid \vec{x}) = \frac{e^{-\vec{w}_j \cdot \vec{x}}}{\sum_{j'=1}^{K} e^{-\vec{w}_{j'} \cdot \vec{x}}}$$

πιθανότητα να ανήκει στην c_j

διαφορετικό διάνυσμα βαρών ανά κατηγορία

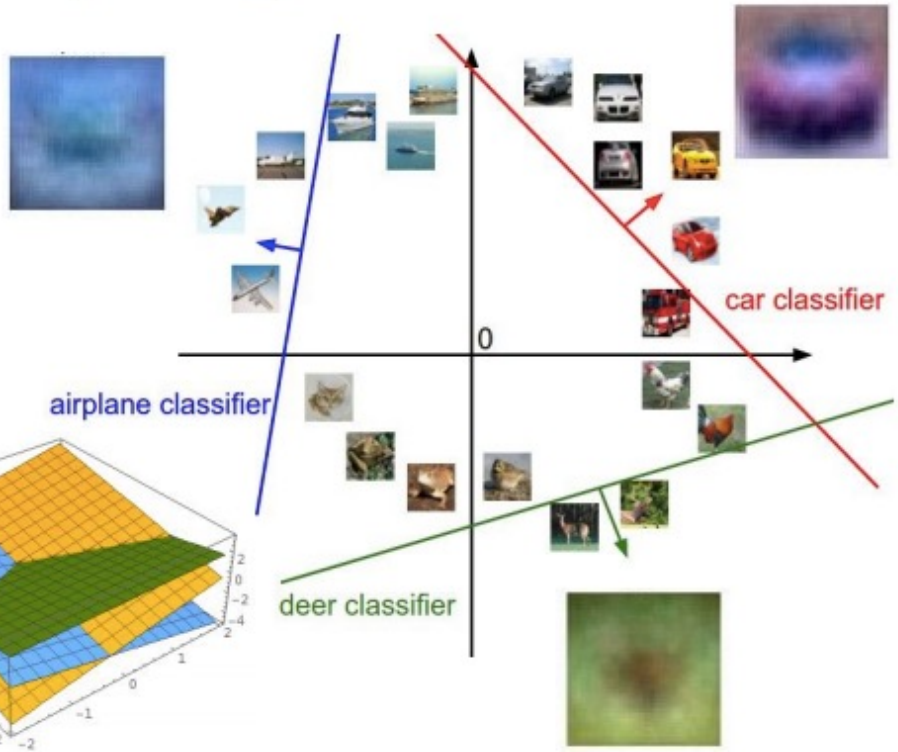
σταθερά κανονικοποίησης

  - Άλλη εξήγηση: υπολογίζουμε την **ομοιότητα** $\boldsymbol{z_j = \vec{w}_j \cdot \vec{x}}$ του $\vec{x}$ με τον **εκπρόσωπο** $\vec{w}_j$ κάθε **κατηγορίας** c_j και εφαρμόζουμε **softmax** $\left(\text{δηλ. } \frac{\exp(z_j)}{\sum_{j'} \exp(z_{j'})}\right)$ στις ομοιότητες z_j για να γίνουν **πιθανότητες** με άθροισμα 1.

- Εκπαιδεύουμε πάλι **μεγιστοποιώντας** τη (δεσμευμένη) **πιθανοφάνεια** των παραδειγμάτων εκπαίδευσης.

Πολυωνυμική λογιστική παλινδρόμηση

Interpreting a Linear Classifier



$$f(x,W) = Wx + b$$

Array of **32x32x3** numbers
(3072 numbers total)

Plot created using Wolfram Cloud

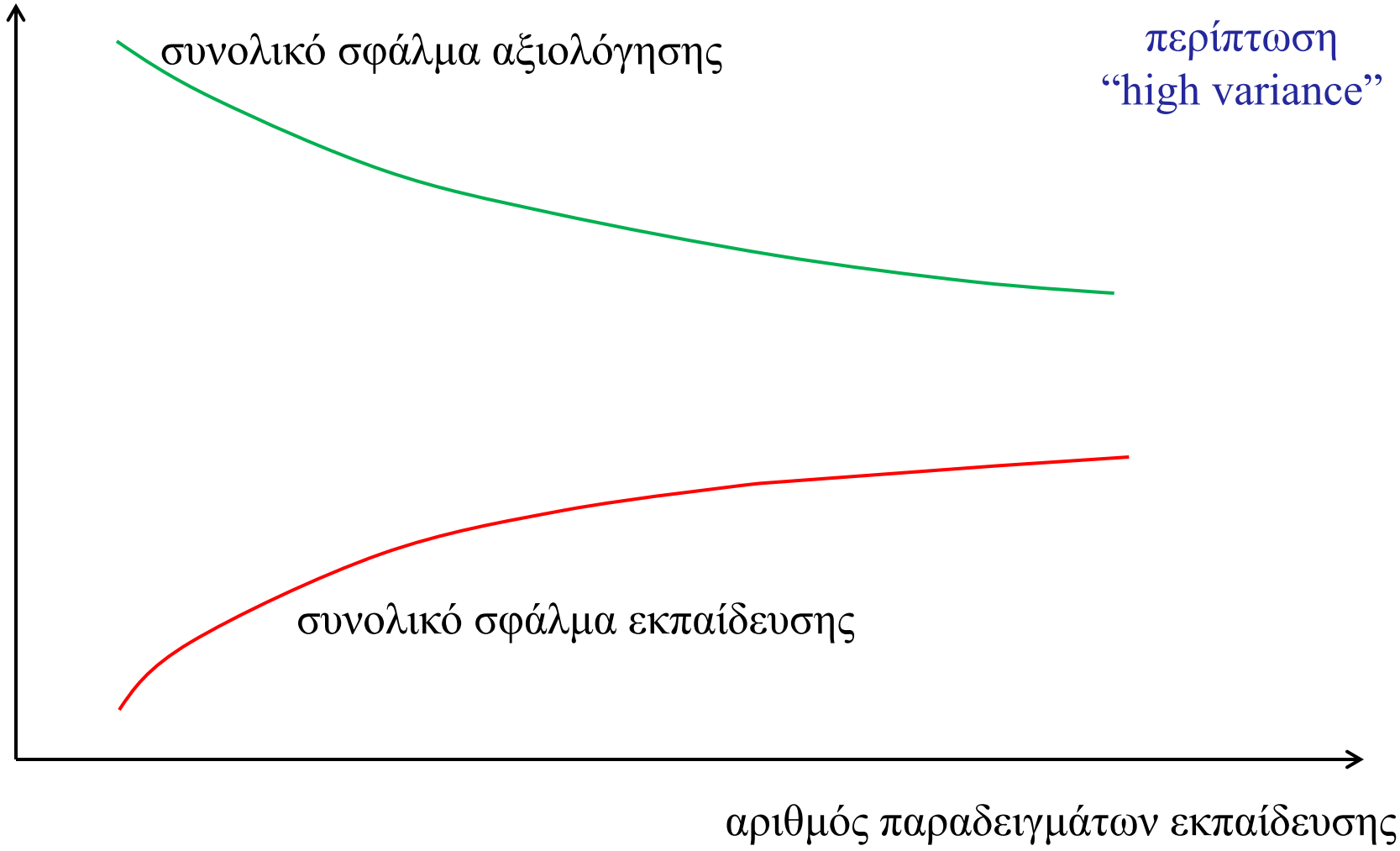
Cat image by Nikita is licensed under CC-BY 2.0

From Stanford's course "Convolutional Neural Networks for Visual Recognition", F.F. Li, J. Johnson, S. Yeung, 2017. http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture2.pdf

Συμβουλές χρήσης επιβλεπόμενης ΜΜ (βασισμένες σε συμβουλές του A. Ng)

- Στους περισσότερους αλγορίθμους **επιβλεπόμενης μηχανικής μάθησης** το συνολικό **σφάλμα στα δεδομένα εκπαίδευσης** είναι **χαμηλότερο** από το συνολικό σφάλμα στα δεδομένα αξιολόγησης.
  - Σφάλμα εκπαίδευσης: Πόσο καλά τα πάμε στα **ίδια δεδομένα** που χρησιμοποιήσαμε για εκπαίδευση.
  - Σφάλμα αξιολόγησης: Πόσο καλά τα πάμε σε **διαφορετικά δεδομένα** από εκείνα που χρησιμοποιήσαμε για εκπαίδευση.
- Το σφάλμα εκπαίδευσης συχνά είναι μια χρήσιμη ένδειξη του **πόσο καλά μπορούμε να ελπίζουμε ότι θα τα πάμε** κατά την αξιολόγηση.
- Παραστάσεις των δύο ειδών σφαλμάτων συχνά βοηθούν να **διαγνώσουμε** τι δεν πάει καλά με το σύστημά μας.

Διαγνωστικοί έλεγχοι: **υπερ-εφαρμογή**



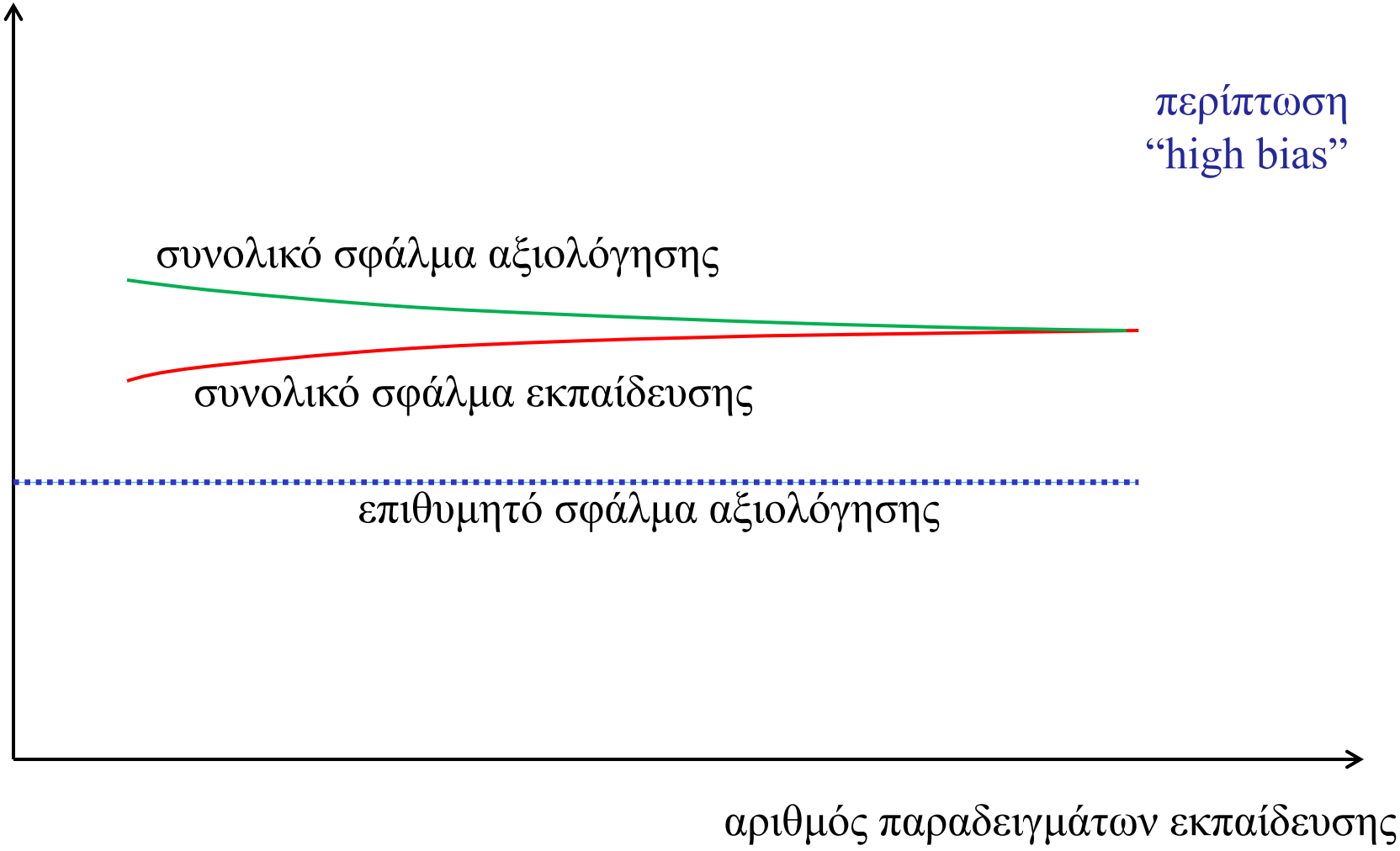
Διαγνωστικοί έλεγχοι: **υπερ-εφαρμογή**

- Αν **παρατηρούμε** τα εξής:
  - Το συνολικό σφάλμα εκπαίδευσης αυξάνεται (**χειροτερεύει**) **απότομα** όσο προσθέτουμε παραδείγματα εκπαίδευσης.
  - Το συνολικό σφάλμα αξιολόγησης μειώνεται (**βελτιώνεται**) **απότομα** όσο προσθέτουμε παραδείγματα εκπαίδευσης.
  - **Κυρίως**: υπάρχει **μεγάλη διαφορά** μεταξύ των δύο σφαλμάτων.
- Μπορεί το σύστημα να πάσχει από **υπερ-εφαρμογή**:
  - Τα πηγαίνει πολύ καλύτερα στα δεδομένα εκπαίδευσης από ό,τι στα δεδομένα αξιολόγησης, γιατί μαθαίνει **ιδιαιτερότητες** των **παραδειγμάτων αξιολόγησης.**
  - Ευκολότερο να συμβεί με **λίγα δεδομένα εκπαίδευσης**.
  - Όσο **αυξάνονται** τα **δεδομένα εκπαίδευσης**, τόσο **δυσκολότερο** γίνεται να μάθει **ιδιαιτερότητές** τους. Πετυχαίνει καλύτερη γενίκευση, οπότε τα πηγαίνει και **καλύτερα** στα δεδομένα αξιολόγησης.

Διαγνωστικοί έλεγχοι: **υπερ-εφαρμογή**

- Τι **μπορεί** να βοηθήσει:
  - **Περισσότερα δεδομένα εκπαίδευσης**.
  - **Λιγότερες** (καλύτερες) **ιδιότητες** (επιλογή ιδιοτήτων, SVD).
  - **Μεγαλύτερο *λ*** στη λογιστική παλινδρόμηση.
  - **Απλούστερο μοντέλο υποθέσεων** (π.χ. γραμμικές αντί για πολυωνυμικές υποθέσεις υψηλότερου βαθμού ή αντί για μη παραμετρικό μοντέλο όπως ο *k*-NN).
- Τι **δεν** θα βοηθήσει μάλλον:
  - **Περισσότερες ιδιότητες**.
  - **Πιο περίπλοκο μοντέλο** (π.χ. πιο πολλά επίπεδα ή περισσότεροι νευρώνες ανά επίπεδο σε ένα νευρωνικό δίκτυο).
  - **Περισσότερα δέντρα στα Τυχαία Δάση**.

Διαγνωστικοί έλεγχοι: πολύ **περιορισμένος χώρος αναζήτησης**



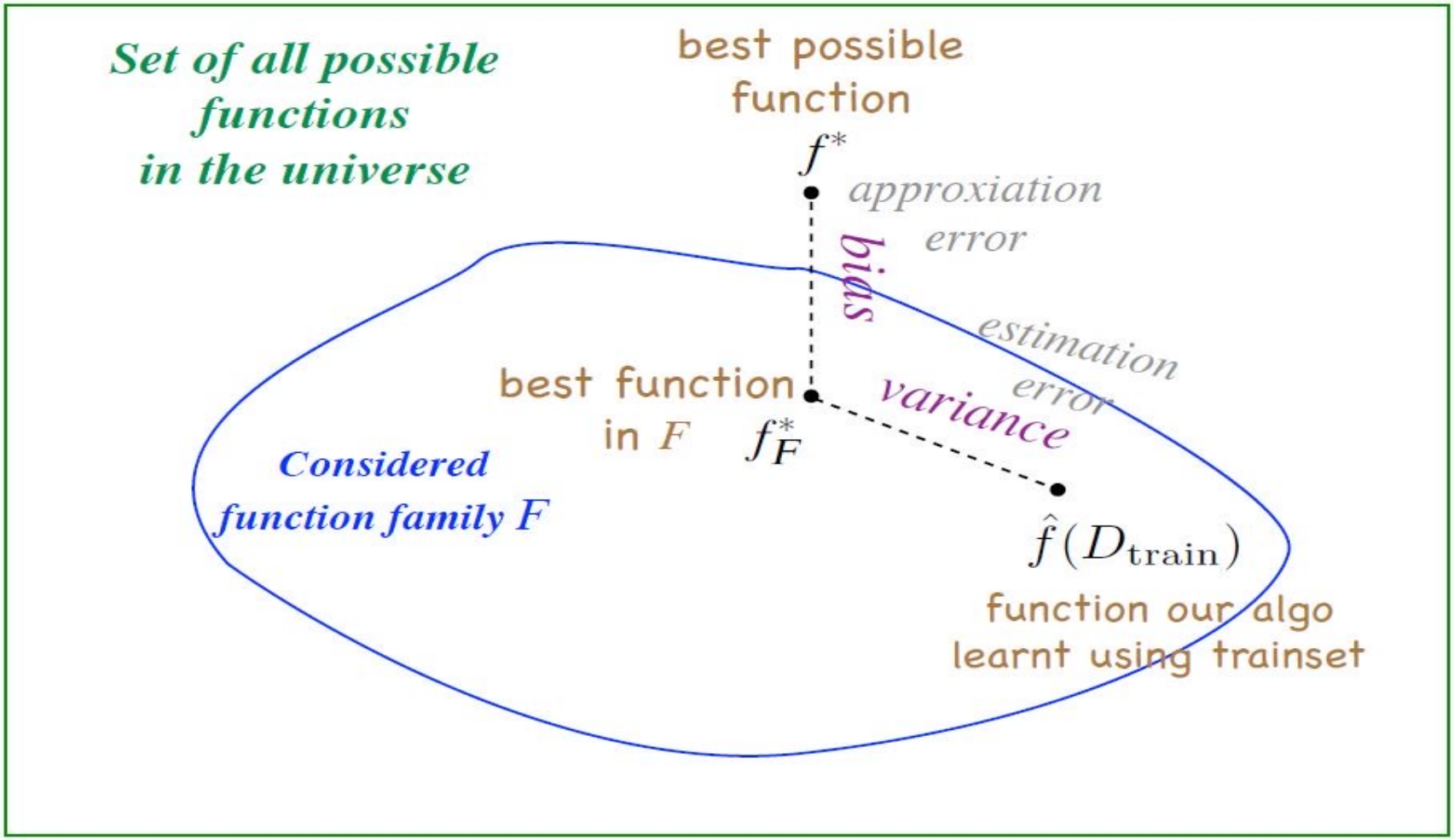
Διαγνωστικοί έλεγχοι: **περιορισμένος χώρος αναζήτησης**

- Αν **παρατηρούμε** τα εξής:
  - Το συνολικό σφάλμα εκπαίδευσης **αυξάνεται (χειροτερεύει) πολύ λίγο** όσο προσθέτουμε παραδείγματα εκπαίδευσης.
  - Το συνολικό σφάλμα αξιολόγησης **μειώνεται (βελτιώνεται) πολύ λίγο** όσο προσθέτουμε παραδείγματα εκπαίδευσης.
  - **Κυρίως**: υπάρχει πολύ **μικρή διαφορά** μεταξύ των δύο σφαλμάτων (και δεν έχουμε φτάσει στο επιθυμητό επίπεδο σφάλματος).
- Ίσως ο **χώρος αναζήτησης** είναι υπερβολικά **περιορισμένος**:
  - Το σύστημα ίσως **δεν μπορεί να μάθει** αυτό που θέλουμε, γιατί δεν περιλαμβάνεται στο χώρο αναζήτησης.
  - Οι **υποθέσεις** του χώρου ίσως είναι **υπερβολικά απλοϊκές**, για να γενικεύσουν τα δεδομένα εκπαίδευσης.

Διαγνωστικοί έλεγχοι: **περιορισμένος χώρος αναζήτησης**

- Τι **μπορεί να βοηθήσει**:
  - **Περισσότερες ιδιότητες** (π.χ. νέες πληροφορίες ή προσθήκη συνδυασμών ιδιοτήτων, όπως λογικό ΚΑΙ ζευγών ιδιοτήτων στη λογιστική παλινδρόμηση, που δεν μαθαίνει από μόνη της τέτοιους συνδυασμούς).
  - **Πιο περίπλοκο μοντέλο υποθέσεων** (π.χ. περισσότερα επίπεδα ή περισσότεροι νευρώνες ανά επίπεδο σε ένα νευρωνικό δίκτυο).
  - **Περισσότερα δέντρα** στα **Τυχαία Δάση**.
  - **Μικρότερο *λ*** στη λογιστική παλινδρόμηση.
- Τι **δεν θα βοηθήσει** μάλλον:
  - **Περισσότερα δεδομένα εκπαίδευσης**.
  - **Λιγότερες ιδιότητες** (π.χ. με επιλογή ιδιοτήτων, SVD).

Decomposing the generalization error



Από την παρουσίαση «Introduction to Machine Learning» του P. Vincent στο Deep Learning Summer School 2015 (http://videolectures.net/deeplearning2015_montreal/).

Evaluating classifiers

- **Accuracy** (correct decisions/total decisions) is **not always** a **good evaluation measure**!
  - If we have two classes and one is much more frequent (e.g., 80% of instances), a **majority classifier** that always classifies in the most frequent class will have an accuracy of 80%!
- **Precision of a class**:
  - **How many** of the instances **classified in the class** (true positives + false positives) **are true members** of the class (true positives).

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

- **Recall of a class**:
  - **How many** of the **true members** of a class (true positives + false negatives) are **classified in the class** (true positives).

Evaluating classifiers – continued

- **F-measure**: $$F_\beta = \frac{(\beta^2 + 1) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$$
  - **Combination** of **precision** and **recall** (weighted harmonic mean).
  - For $\boldsymbol{\beta = 1}$, **equal importance** to **precision** and **recall**. (But the harmonic mean is closer to the min of the two values than the arithmetic mean.)
- **Averaging precision or recall** over ***n*** **classes**:
  - **Macro-averaging** (equal weight assigned to all classes):

$$\text{MacroPrecision} = \frac{1}{n}\sum_{i=1}^{n} \text{Precision}_i \qquad \text{MacroRecall} = \frac{1}{n}\sum_{i=1}^{n} \text{Recall}_i$$

  - **Micro-averaging** (frequent classes treated as more important):

$$\text{MicroPrecision} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + FP_i} \qquad \text{MicroRecall} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + FN_i}$$

Precision-recall diagrams

- In many algorithms, we can **opt for higher precision** at the expense of **lower recall**, or **vice versa** by **tuning** a **threshold**.
  - In Naive Bayes: $h(\vec{x}) = 1$ iff $P(C = 1|\vec{x}) > \boldsymbol{t}$
  - For **different values** of the **threshold $\boldsymbol{t}$**, we obtain **different pairs of precision-recall scores** (on test data).
  - The **larger** the **area under the curve** (**AUC of Precision-Recall** curve, a.k.a. **Average Precision**) the **better** the system. (AP is slightly different in IR.)
  - For **multiple classes**, we can **average AP** over classes, obtaining **Mean Average Precision** (**MAP**).

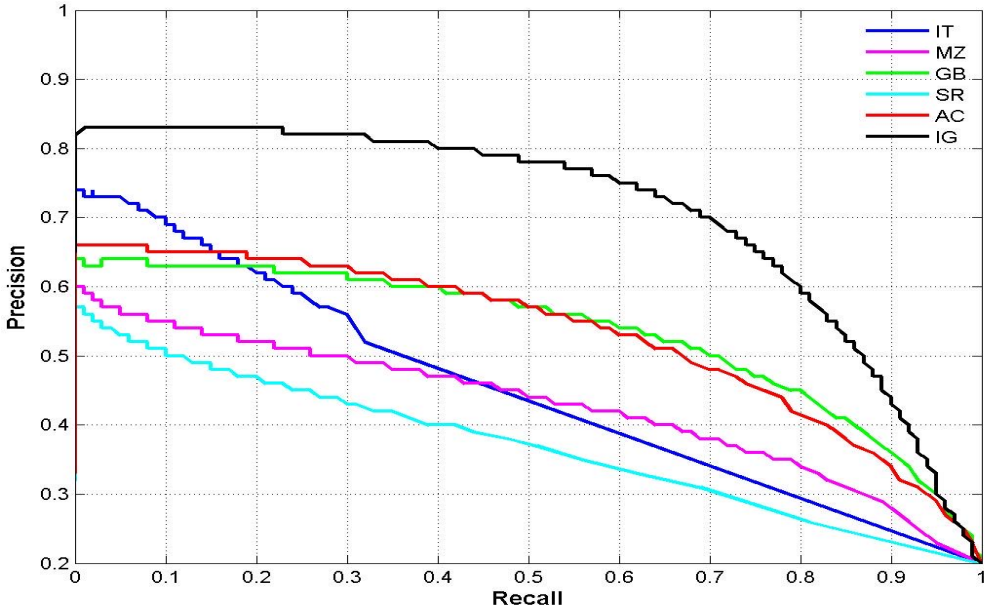


Image source: http://ivrgwww.epfl.ch/supplementary_material/RK_CVPR09/Images/prcurve_1000.jpg

ROC curves

- Instead of Precision-Recall curves, it is also common to plot **Receiver Operating Characteristic** (ROC) curves.
  - **True Positive Rate** = $\frac{TP}{TP+FN}$ = **Sensitivity** = Recall of positive class
  - **False Positive Rate** = $\frac{FP}{FP+TN}$ = 1 – $\frac{TN}{TN+FP}$ = **1 – Specificity** = 1 – Recall of negative class
  - The **larger** the **AUC** (of ROC curve) the **better** the system.

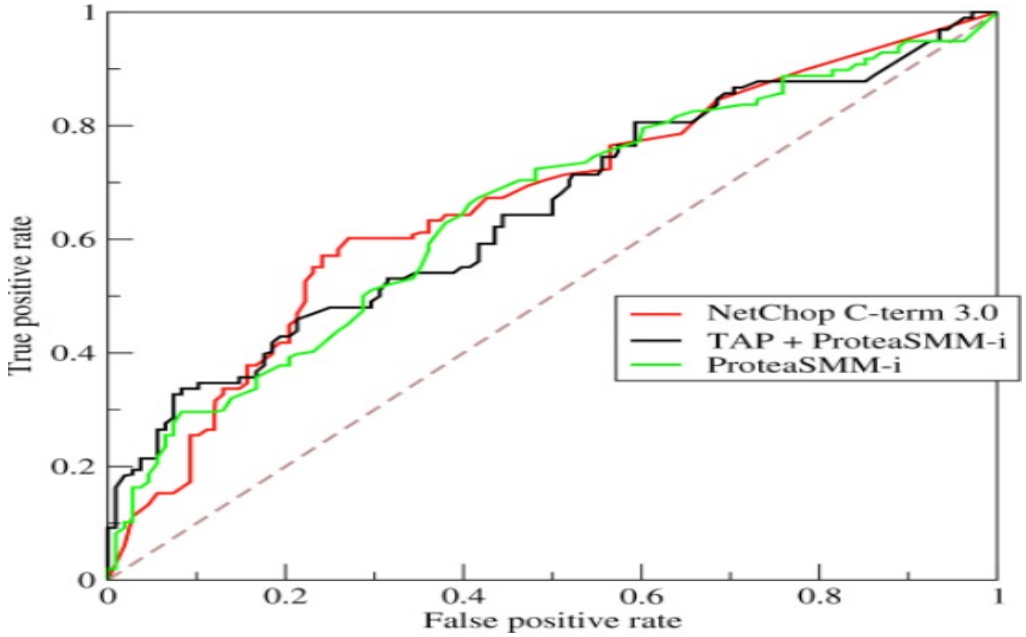


Image source: https://en.wikipedia.org/wiki/Receiver_operating_characteristic#/media/File:Roccurves.png

Βιβλιογραφία

- Russel & Norvig (4η έκδοση): ενότητες 19.6.4, 19.6.5, 19.9.3.
  - Όσοι ενδιαφέρονται μπορούν να διαβάσουν προαιρετικά και τις υπόλοιπες ενότητες του κεφαλαίου 19.
- Βλαχάβας κ.ά: ενότητες 18.3.3, 18.12.
- Συμβουλευτείτε τις σημειώσεις «Linear regression, classification and logistic regression, generalized linear models» του A. Ng.
  - Βλ. https://sgfin.github.io/files/notes/CS229_Lecture_Notes.pdf, σελ. 1–7, 16–19.
  - Όσοι ενδιαφέρονται μπορούν να διαβάσουν προαιρετικά (εκτός εξεταστέας ύλης) και τις υπόλοιπες ενότητες. Το Perceptron της ενότητας 6 θα το συναντήσουμε και στην επόμενη διάλεξη.

Βιβλιογραφία – συνέχεια

- Οι ταξινομητές λογιστικής παλινδρόμησης περιγράφονται και σε πρόσθετο (ηλεκτρονικό, δωρεάν) κεφάλαιο του βιβλίου «Machine Learning» του T. Mitchell.
  - Βλ. http://www.cs.cmu.edu/~tom/NewChapters.html.
  - Βλ. εισαγωγή ενότητας 3 και υπο-ενότητες 3.2 και 3.3.

- Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines, SVM) είναι άλλη μια σημαντική μέθοδος επιβλεπόμενης μάθησης.
  - Περιγράφονται στην ενότητα 19.7.5 των Russel & Norvig (4η έκδοση) και στην ενότητα 18.9 των Βλαχάβα κ.ά.
  - Περιγράφονται επίσης στο κεφάλαιο 15 του βιβλίου «An Introduction to Information Retrieval» των C.D. Manning, P. Raghavan και H. Schütze, το οποίο διατίθεται ελεύθερα (βλ. http://www-nlp.stanford.edu/IR-book/).