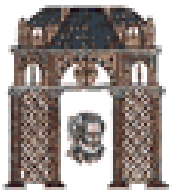
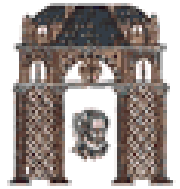


Αναγνώριση Προτύπων

Εκτίμηση Παραμέτρων
(Parameter Estimation)

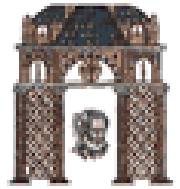
Χριστόδουλος Χαμζάς

Τα περιεχόμενα των παρουσιάσεων προέρχονται κυρίως από τις παρουσιάσεις του αντίστοιχου διδακτέου μαθήματος του καθ. Παναγιώτη Τσακαλίδη, Τμ. Επιστήμης Υπολογιστών, Παν. Κρήτης. Το πρωτογενές υλικό βρίσκεται στην σελίδα http://www.csd.uoc.gr/~hy473/ *και βασίζεται στο βιβλίο: "Pattern Classification", R.O. Duda, P.E. Hart, D.G. Stork, Wiley, 2nd Ed., 2001*

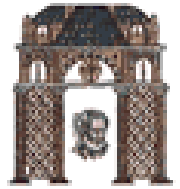


Εκτίμηση Παραμέτρων

- Ο Μπεϋζιανός ταξινομητής δεν μπορεί να χρησιμοποιηθεί όταν δεν είναι γνωστές οι συναρτήσεις πυκνότητας πιθανότητας $p(\boldsymbol{x}/\omega_i)$ & $P(\omega_i)$.
- Οι κατανομές μπορούν να εκτιμηθούν εάν υπάρχουν επαρκή δεδομένα→Δύσκολο!
- Εάν είναι γνωστή η μορφή της κατανομής, για παράδειγμα κανονική, αλλά όχι οι παράμετροι της, π.χ. η μέση τιμή και η διασπορά της, το πρόβλημα ανάγεται σε αυτό της εκτίμησης των παραμέτρων.



- Υπάρχουν επίσης δύο τεχνικές μάθησης
- 1. Εκπαίδευση με επίβλεψη (supervised learning)
  - Υπάρχει αρχικό δείγμα στο οποίο γνωρίζουμε σε ποια κατηγορία ανήκει το καθένα (π.χ. έχουμε ένα δείγμα από χαρακτηριστικά στο οποίο ξέρουμε ποια είναι από λαυράκια και ποια είναι από σολωμούς)
- 2. Εκπαίδευση δίχως επίβλεψη (unsupervised learning)
  - Δεν υπάρχει αρχικό δείγμα εκπαίδευσης απλά ξέρουμε ότι τα δείγματα μας προέρχονται από γνωστές κατηγορίες (π.χ. έχουμε ένα δείγμα από χαρακτηριστικά και απλά ξέρουμε ότι προέρχονται μόνο από λαυράκια και από σολωμούς)
  - Το πρόβλημα αυτό θα το αφήσουμε για άλλο μάθημα.



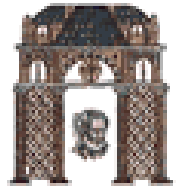
Εκτίμηση Παραμέτρων (2)

- Υπάρχουν δύο βασικές τεχνικές:
  1. Εκτίμηση Μέγιστης Πιθανοφάνειας (Maximum Likelihood Estimation)
  2. Εκτίμηση παραμέτρων κατά Bayes (Bayesian Parameter Estimation)

Τα αποτελέσματα και των 2 τεχνικών είναι περίπου ίδια αλλά η θεωρητική προσέγγιση του προβλήματος είναι διαφορετική

1.Η εκτίμηση μεγίστης πιθανοφάνειας θεωρεί τις τιμές των αγνώστων παραμέτρων «άγνωστες σταθερές» και προσπαθεί να βρεί ποιες είναι αυτές ώστε να μεγιστοποιείται η πιθανότητα να πάρουμε το συγκεκριμένο «δείγμα» τιμών.

2. Η εκτίμηση παραμέτρων κατά Bayes, αντίθετα θεωρεί ότι οί άγνωστοι παράμετροι είναι τυχαίες μεταβλητές με γνωστή από πριν αρχική πυκνότητα κατανομής (prior) και το δείγμα τιμών που έχουμε λάβει μας οδηγεί σε βελτίωση (posterior) της αρχικής μας κατανομής.



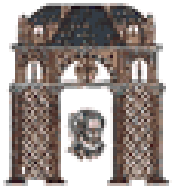
Εκτίμηση Μέγιστης Αληθοφάνειας

- Υποθέτουμε ότι παίρνουμε τυχαία επιλεγμένα δείγματα από μια δεδομένη κατανομή της οποίας οι παράμετροι είναι άγνωστοι. Συμβολίζουμε το σύνολο των παραμέτρων με το διάνυσμα θ.

- Εάν για παράδειγμα γνωρίζουμε ότι η κατανομή είναι κανονική αλλά δεν ξέρουμε τον μέσο και τη διασπορά της, τότε

$$\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mu_1, \ldots, \mu_d, \sigma_1^2, \ldots, \sigma_d^2, \mathrm{cov}(x_m, x_n); m, n = 1, \ldots d; m \succ n)$$

$$d + \frac{d(d+1)}{2} \quad \text{παράμετροι}$$



Πρόβλημα MLE

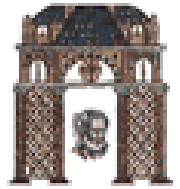
- Για κάθε μία από τις κλάσεις, υπολόγισε το διάνυσμα παραμέτρων θ χρησιμοποιώντας ένα σύνολο δεδομένων εκπαίδευσης $D^n=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ το οποίο έχει n ανεξάρτητα δείγματα (i.i.d):

$$p(D^n \mid \boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

Πιθανοφάνεια του θ αναφορικά με το δείγμα D^n

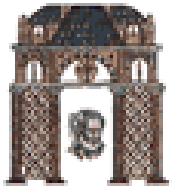
- Η εκτίμηση μεγίστης πιθανοφάνειας του θ είναι εκείνη η τιμή που μεγιστοποιεί την παραπάνω συνάρτηση.

- Διαισθητικά, αντιστοιχεί στην τιμή του $\boldsymbol{\theta}$ που «συμφωνεί» όσο το δυνατό καλύτερα με τα δείγματα.



Παράδειγμα με νόμισμα

- Ρίχνουμε ένα νόμισμα Ν φορές και παρατηρούμε το αποτέλεσμα.
- Ποια είναι η εκτίμηση της πιθανότητας να έχουμε «γράμματα» ?



Συνάρτηση Πιθανοφάνειας

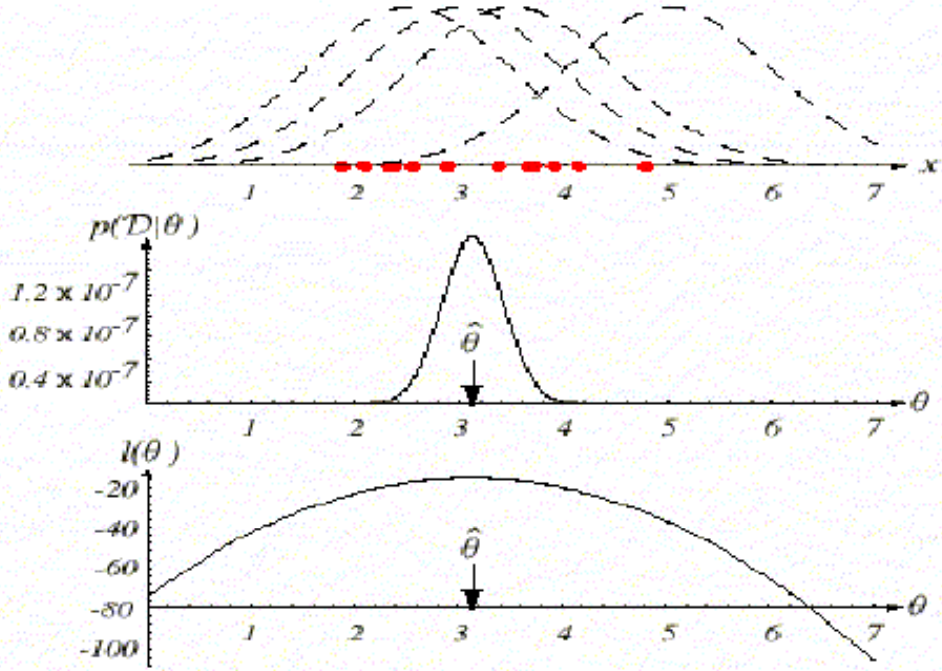
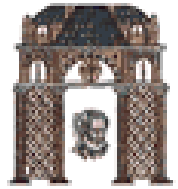


FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$; it also maximizes the logarithm of the likelihood—that is, the log-likelihood $l(\theta)$, shown at the bottom. Note that even though they look similar, the likelihood $p(\mathcal{D}|\theta)$ is shown as a function of θ whereas the conditional density $p(x|\theta)$ is shown as a function of x. Furthermore, as a function of θ, the likelihood $p(\mathcal{D}|\theta)$ is not a probability density function and its area has no significance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Λογάριθμος της Πιθανοφάνειας (Log-Likelihood)

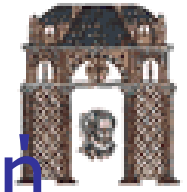


- Το θ που μεγιστοποιεί την συνάρτηση πιθανοφάνειας, μεγιστοποιεί επίσης και τον λογάριθμό της, με τον οποίο μπορούμε πολλές φορές να δουλέψουμε ευκολότερα:

$$l(\boldsymbol{\theta}) \equiv \ln p(D^n \mid \boldsymbol{\theta}) = \sum_{k=1}^{n} \ln p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

Το θ που μεγιστοποιεί αυτή τη συνάρτηση μπορεί να υπολογισθεί θέτοντας την παράγωγο του ως προς θ ίση με το μηδέν, και λύνοντας την εξίσωση ως προς θ. $\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k \mid \boldsymbol{\theta}) = \mathbf{0}$$



Ειδικές Περιπτώσεις – Κανονική κατανομή

- Περίπτωση κανονικής κατανομής με άγνωστο μ
  - $p(\mathbf{x}_\kappa \mid \mu) \sim N(\mu, \Sigma)$

$$\ln p(\mathbf{x}_k \mid \boldsymbol{\mu}) = -\frac{1}{2}\ln\left[(2\pi)^d \left|\boldsymbol{\Sigma}\right|\right] - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

$$\nabla_\theta \ln p(\mathbf{x}_k \mid \boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

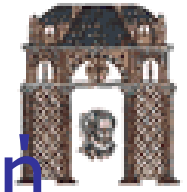
  - Ο εκτιμητής μέγιστης πιθανοφάνειας του $\boldsymbol{\mu}$ ικανοποιεί:

$$\sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}) = 0$$

  - Πολλαπλασιάζοντας από αριστερά με Σ και λύνοντας, παίρνουμε:

$$\hat{\boldsymbol{\mu}}_{ML} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$$

Δηλαδή είναι ο αριθμητικός μέσος όρος των δειγμάτων εκπαίδευσης!



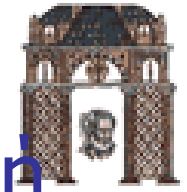
Ειδικές Περιπτώσεις – Κανονική κατανομή

- Περίπτωση κανονικής κατανομής με αγνώστους μ και σ^2:
  - $\boldsymbol{\theta} = (\theta_1, \theta_2) = (\mu, \sigma^2)$

$$l(\boldsymbol{\theta}) = \ln P(x_k \mid \boldsymbol{\theta}) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_{\boldsymbol{\theta}} l = \begin{pmatrix} \frac{\partial}{\partial \theta_1}(\ln P(x_k \mid \boldsymbol{\theta})) \\ \frac{\partial}{\partial \theta_2}(\ln P(x_k \mid \boldsymbol{\theta})) \end{pmatrix} = \mathbf{0}$$

$$\begin{cases} \frac{1}{\theta_2}(x_k - \theta_1) = 0 \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} = 0 \end{cases}$$



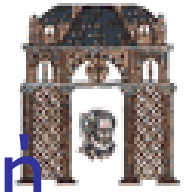
Ειδικές Περιπτώσεις – Κανονική κατανομή

Λαμβάνοντας υπ' όψην όλα τα δείγματα :

$$\begin{cases} \sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 & (1) \\ -\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 & (2) \end{cases}$$

Συνδυάζοντας τις (1) και (2), παίρνουμε:

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \quad ; \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2$$



Ειδικές Περιπτώσεις – Κανονική κατανομή

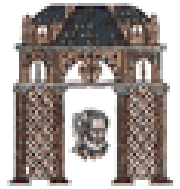
- Bias

  - Ο ML εκτιμητής για τη διασπορά σ^2 μεροληπτεί (biased)

$$E\left[\frac{1}{n}\sum_{k=1}^{n}(x_k-\hat{\mu})^2\right]=\frac{n-1}{n}\sigma^2\neq\sigma^2$$

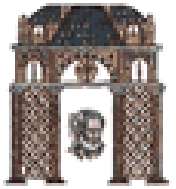
  - Ένας στοιχειώδης αμερόληπτος εκτιμητής για τον πίνακα συνδιασποράς, Σ, είναι:

$$\underbrace{\mathbf{C}=\frac{1}{\mathrm{n}\text{ -}1}\sum_{k=1}^{n}(\mathbf{x}_k-\boldsymbol{\mu})(\mathbf{x}_k-\hat{\boldsymbol{\mu}})^t}_{\textit{Sample}\ \text{covariance matrix}}$$



Εκτίμηση κατά Bayes

- Σε αντίθεση με τον MLE, όπου κάναμε την υπόθεση ότι οι άγνωστες παράμετροι έχουν σταθερές τιμές, ο εκτιμητής κατά Bayes (BE) υποθέτει ότι οι άγνωστες παράμετροι είναι τυχαίες μεταβλητές και ακολουθούν μία εκ των προτέρων γνωστή σ.π.π.
- Επομένως, ο BE υπολογίζει μια κατανομή των τιμών του θ και όχι τις τιμές αυτές καθ' αυτές. Ο BE παρέχει περισσότερη πληροφορία, όμως συχνά είναι δύσκολο να υπολογισθεί.
- Η ύπαρξη μετρήσεων εκπαίδευσης (training data) επιτρέπει την μετατροπή της εκ των προτέρων πληροφορίας σε εκ των υστέρων σ.π.π. → φαινόμενο της εκμάθησης (Bayesian learning) όπου κάθε νέα παρατήρηση οξύνει την εκ των υστέρων σ.π.π.

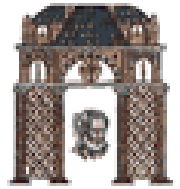


Εκτίμηση κατά Bayes

- Μπεϋζιανή Εκμάθηση για προβλήματα ταξινόμησης προτύπων.

  - Ο υπολογισμός των εκ των υστέρων σ.π.π. αποτελεί τη βάση της Μπεϋζιανής ταξινόμησης.
  - Στόχος: Υπολογισμός των $P(\omega_i \mid \mathbf{x}, D)$ δεδομένου του συνόλου δειγμάτων εκπαίδευσης $D=\{D_1,\ldots,D_c\}$, όπου τα δείγματα στο σύνολο D_j αντιστοιχούν στην κλάση $j, j=1,\ldots,c$.
  - Για κάθε νέο, αταξινόμητο δείγμα, $\mathbf{x}$, ο κανόνας του Bayes δίνει:

$$P(\omega_i \mid \mathbf{x}, D) = \frac{p(\mathbf{x} \mid \omega_i, D)P(\omega_i \mid D)}{\sum_{j=1}^{c} p(\mathbf{x} \mid \omega_j, D)P(\omega_j \mid D)}$$

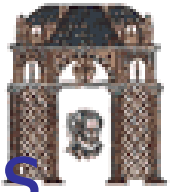


Εκτίμηση κατά Bayes

- Υποθέτουμε ότι
  - Οι εκ των προτέρων $P(\omega_i)$ είναι γνωστές, οπότε $P(\omega_i/D) = P(\omega_i)$.
  - Μόνο τα δείγματα του συνόλου D_i έχουν πληροφορία για τη σ.π.π. $p(\mathbf{x}/\omega_i, D_i)$ → προκύπτουν c ανεξάρτητα προβλήματα εκτίμησης των $p(\mathbf{x}/\omega_i, D_i)$, η οποία μπορεί να γραφεί και ως $p(\mathbf{x}/D_i)$.

$$P(\omega_i \mid \mathbf{x}, D_i) = \frac{p(\mathbf{x} \mid \omega_i, D_i) P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x} \mid \omega_j, D_j) P(\omega_j)}$$

Αυτή τη συνάρτηση θέλουμε να εκτιμήσουμε (γράφεται και ως $p(\mathbf{x}/Di)$)



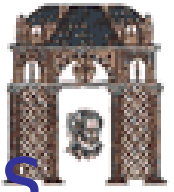
Γενική μεθοδολογία εκτίμησης κατά Bayes

- Για κάθε κλάση, γνωρίζουμε τη μορφή της σ.π.π. $p(\mathbf{x}|\boldsymbol{\theta})$, αλλά η τιμή του διανύσματος παραμέτρων $\boldsymbol{\theta}$ είναι άγνωστη.
- Έχουμε κάποια αρχική γνώση για το $\boldsymbol{\theta}$ με τη μορφή της *εκ των προτέρων (a priori)* σ.π.π. $p(\boldsymbol{\theta})$.
- Για κάθε κλάση, διαθέτουμε ένα σύνολο $D^n=\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ από n στατιστικώς ανεξάρτητα δείγματα. Τότε:

$$p(\mathbf{x}/D)=\int_{\boldsymbol{\theta}} p(\mathbf{x}/\boldsymbol{\theta})\,p(\boldsymbol{\theta}/D)\,d\boldsymbol{\theta}$$

Σχέση κλειδί: Συνδέει την δεσμευμένη σ.π.π. $p(\mathbf{x}/D)$ με την εκ των υστέρων σ.π.π. $p(\boldsymbol{\theta}/D)$ του διανύσματος των παραμέτρων. Δηλώνει ότι η $p(\mathbf{x}/D)$ είναι ένας γραμμικός συνδιασμός των $p(\mathbf{x}/\boldsymbol{\theta})$ με βάρη τις $p(\boldsymbol{\theta}/D)$.

Άν η $p(\boldsymbol{\theta}/D)$ έχει ένα απότομο μοναδικό μέγιστο στο $\boldsymbol{\theta}^*$ τότε $p(\mathbf{x}/D)\approx p(\mathbf{x}/\boldsymbol{\theta}^*)$



Γενική μεθοδολογία εκτίμησης κατά Bayes

- Για τον υπολογισμό της $p(\boldsymbol{\theta}/D)$, έχουμε ότι:

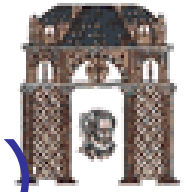
$$p(\boldsymbol{\theta}/D)=\frac{p(D/\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(D/\boldsymbol{\theta})\,p(\boldsymbol{\theta})\,d\boldsymbol{\theta}}$$

- Λόγω της ανεξαρτησίας των δειγμάτων εκπαίδευσης, $p(D/\boldsymbol{\theta})=\prod_{k=1}^{n} p(\mathbf{x}_k/\boldsymbol{\theta})$

- Αν η $p(D/\boldsymbol{\theta})$ είναι επικεντρωμένη γύρω από το $\boldsymbol{\theta}^*$ με μεγάλη κορυφή σε αυτό το σημείο, και αν η $p(\boldsymbol{\theta}^*)$ δεν είναι 0, τότε και η $p(\boldsymbol{\theta}/D)$ έχει μεγάλη κορυφή στο $\boldsymbol{\theta}^*$, και επομένως θα είναι

$$p(\mathbf{x}/D)\approx p(\mathbf{x}/\boldsymbol{\theta}^*)$$

- Αλλά το σημείο $\boldsymbol{\theta}^*$, όπως περιγράφεται παραπάνω, είναι ο εκτιμητής μέγιστης πιθανοφάνειας του $\boldsymbol{\theta}$!!



Μπεϋζιανή εκμάθηση (Bayesian Learning)

- Ένα θέμα που ενδιαφέρει είναι αυτό του υπολογισμού και της σύγκλισης της ακολουθίας των σ.π.π. $p(\boldsymbol{\theta}/D^n)$, όπου επαναφέραμε τον δείκτη n του αριθμού των δειγμάτων εκπαίδευσης στο σύνολο D^n.

$$p(D^n/\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k/\boldsymbol{\theta}) = p(\mathbf{x}_n/\boldsymbol{\theta})\, p(D^{n-1}/\boldsymbol{\theta})$$

- Επομένως,

$$p(\boldsymbol{\theta}/D^n) = \frac{p(\mathbf{x}_n/\boldsymbol{\theta})\, p(\boldsymbol{\theta}/D^{n-1})}{\int_{\boldsymbol{\theta}} p(\mathbf{x}_n/\boldsymbol{\theta})\, p(\boldsymbol{\theta}/D^{n-1})\, d\boldsymbol{\theta}}$$

$$p(\boldsymbol{\theta}/D^0) = p(\boldsymbol{\theta})$$

- Η παραπάνω σχέση δημιουργεί μία ακολουθία σ.π.π. $p(\boldsymbol{\theta}/\mathbf{x}_1)$, $p(\boldsymbol{\theta}/\mathbf{x}_1, \mathbf{x}_2)$, …, $p(\boldsymbol{\theta}/\mathbf{x}_1, \ldots, \mathbf{x}_n)$ → Αναδρομική Μπεϋζιανή μεθοδολογία εκμάθησης, Bayesian recursive (or incremental) learning.

Μπεϋζιανή εκμάθηση–Κανονική Κατανομή

- Πρόβλημα: Υπολογισμός των σ.π.π. $p(\boldsymbol{\theta}/D^n)$ και $p(\boldsymbol{x}/D^n)$ όταν υποθέτουμε ότι $p(\boldsymbol{x}/\theta)=p(\boldsymbol{x}/\mu)=N(\mu,\sigma^2)$ (δηλ. $\theta=\mu$) και $p(\mu)=N(\mu_0,\sigma_0^2)$. Το μ_0 είναι η εκ των προτέρων καλύτερη γνώση μας για το μ και το σ_0^2 δηλώνει την αβεβαιότητά μας.
- Προκύπτει ότι $p(\mu/D^n) = N(\mu_n,\sigma_n^2)$, όπου:
- Το μ_n αντιπροσωπεύει την καλύτερη γνώση μας για το μ μετά την παρατήρηση n δειγμάτων εκπαίδευσης και το σ_n^2 μετράει την αβεβαιότητά μας.
- Επίσης, $p(\boldsymbol{x}/D^n)= N(\mu_n, \sigma^2+\sigma_n^2)$.

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0,$$

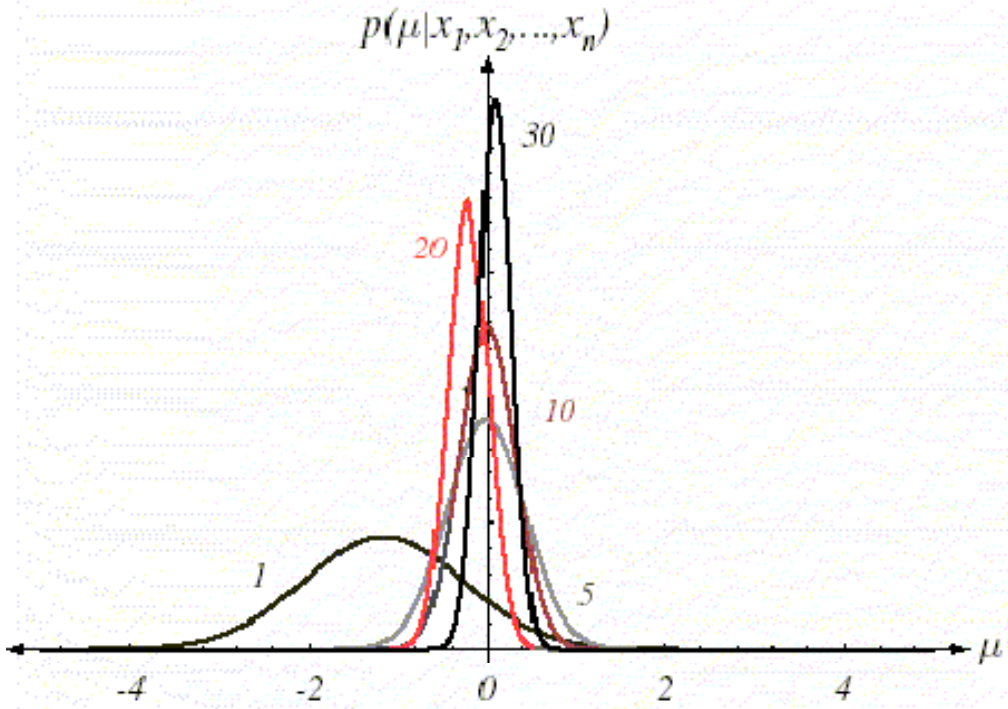
$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2},$$

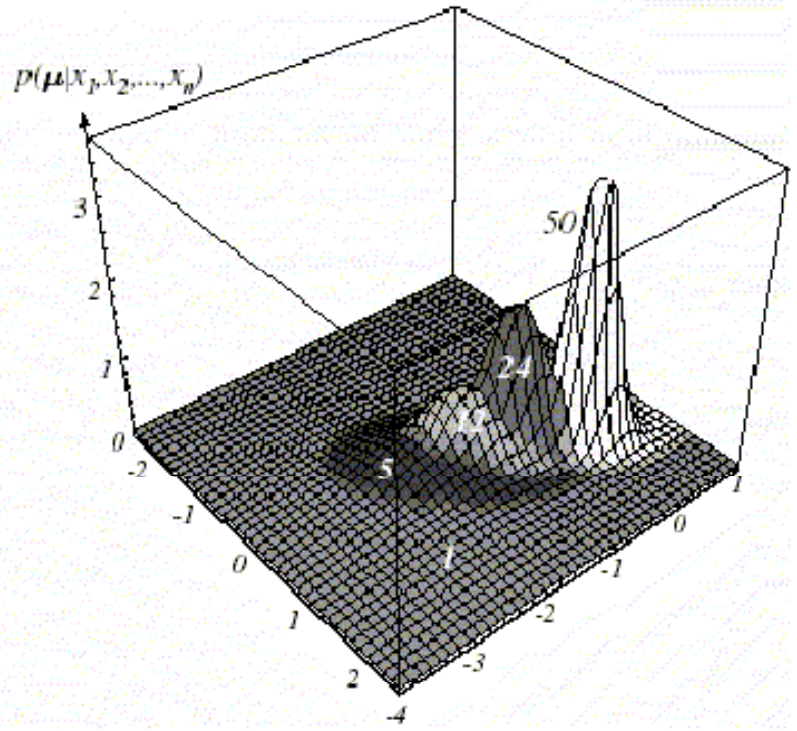
$$\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k$$

✓ Καθώς $\sigma_0^2 \to inf$, έχουμε $\mu_n=\hat{\mu}_n$ για κάθε n, δηλαδή επιστρέφουμε στην εκτίμηση μέγιστης πιθανοφάνειας.

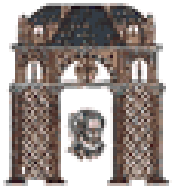
✓ Καθώς $n \to inf$, έχουμε $\sigma_n^2 \approx \sigma^2/n$ δηλαδή για αρκετά μεγάλο αριθμό δειγμάτων εκπαίδευσης, η ακρίβεια της εκτίμησης του μ δεν εξαρτάται από την αβεβαιότητα της εκ των προτέρων γνώσης μας, σ_0^2.

Μπεϋζιανή εκμάθηση–Κανονική Κατανομή





Καθώς $n \to inf$, η εκ των υστέρων σ.π.π. $p(\mu|D^n)$ γίνεται όλο και περισσότερο συγκεντρωμένη γύρω από το μέσο της. Το φαινόμενο αυτό ονομάζεται *Μπεϋζιανή εκμάθηση (Bayesian learning)*.



Παράδειγμα : Εκτίμηση Παραμέτρων

Πετάμε ένα νόμισμα N φορές και παρατηρούμε k φορές κεφάλι. Ποια είναι η πιθανότητα **θ** να φέρουμε κεφάλι;

Μεγίστη Πιθανοφάνεια

Για να πάρουμε την συγκεκριμένη σειρά δεδομένων με κάποιο p έχουμε πιθανότητα $P(D^N \mid \theta) = \theta^k (1-\theta)^{N-k}$

και συνεπώς $P(D^N \mid \theta) = \theta^k (1-\theta)^{N-k}$, $l(\theta) = \ln\left(P(D^N \mid \theta)\right) = k\ln(\theta) + (N-k)\ln(1-\theta)$, $\frac{\partial(l(\theta))}{\partial\theta} = \frac{k}{\theta} - \frac{N-k}{1-\theta} = 0$

Άρα $\hat{\theta} = \frac{k}{N}$

Εκτιμητής Bayes (Αναδρομική Εκτίμηση)

$$p(\boldsymbol{\theta} \mid D^n) = \frac{p(x_n \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D^{n-1})}{\int p(x_n \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D^{n-1})\, d\boldsymbol{\theta}}, \quad p(\boldsymbol{\theta} \mid D^0) \text{ γνωστό (apriori)}$$

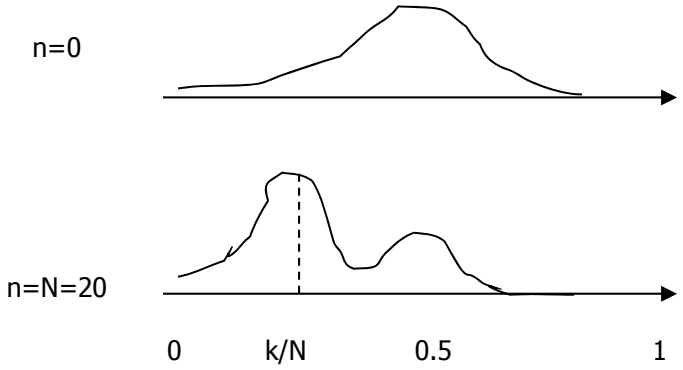
$$\text{ομως} \quad p(x_n \mid \boldsymbol{\theta}) = \begin{cases} \boldsymbol{\theta} & \text{εάν κεφάλι} \\ (1-\boldsymbol{\theta}) & \text{εάν γράμματα} \end{cases}, \text{ και συνεπώς έχουμε}$$

$$p(\boldsymbol{\theta} \mid D^N) = \frac{\boldsymbol{\theta}^k (1-\boldsymbol{\theta})^{N-k}\, p(\boldsymbol{\theta} \mid D^0)}{\int \boldsymbol{\theta}^k (1-\boldsymbol{\theta})^{N-k}\, p(\boldsymbol{\theta} \mid D^0)\, d\boldsymbol{\theta}}$$

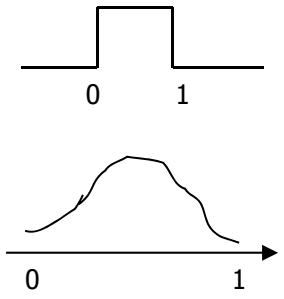
Εκ των προτέρων (a-priori) Πιθανότητα

$p(\boldsymbol{\theta} \mid D^0) = 1$ γιά $0 \le \boldsymbol{\theta} \le 1$ ή

$p(\boldsymbol{\theta} \mid D^0) = Ae^{-\frac{(\boldsymbol{\theta}-0.5)^2}{0,08}} \left(u(\boldsymbol{\theta}) - u(\boldsymbol{\theta}-1)\right)$



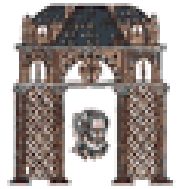
p(θ|D^{20}) για K=4, N=20



Πλήρης άγνοια, ομοιόμορφη κατανομή στο [0,1]

Ή

Αναμένεται να είναι δίκαιο νόμισμα, Κανονική στο [0,1]



Παράδειγμα : Εκτίμηση Παραμέτρων

Έχουμε δεδομένα από μία ομοιόμορφη κατανομή U(0,θ), δηλαδή, p(x|θ)= U(0,θ)=1/θ για 0≤x≤ θ. Τα δεδομένα που έχουμε είναι D={4,7,2,8} και 0≤θ≤10 θέλουμε να εκτιμήσουμε πρώτα το θ και μετά το p(x|θ). Εκτίμηση του θ

Μεγίστη Πιθανοφάνεια

Για να πάρουμε την συγκεκριμένη σειρά δεδομένων με κάποιο θ έχουμε πιθανότητα $p(D^n \mid \theta) = \prod_{k=1}^{n} p(x_k \mid \theta) = (1/\theta)^n \quad 8 \le \theta \le 10$

και συνεπώς $\hat{\theta} = \max\{D_n\} = 8$

Εκτιμητής Bayes (Αναδρομική Εκτίμηση)

$$p(\theta \mid D^n) = \frac{p(x_n \mid \theta) p(\theta \mid D^{n-1})}{\int_{\theta} p(x_n \mid \theta) p(\theta \mid D^{n-1}) d\theta}, \quad p(\theta \mid D^0) = p(\theta) = 0.1, \quad 0 \le \theta \le 10$$

$$p(x \mid \theta) \approx u(0,\theta) = \begin{cases} 1/\theta & 0 \le x \le 1 \\ 0 & \text{αλλού} \end{cases}$$

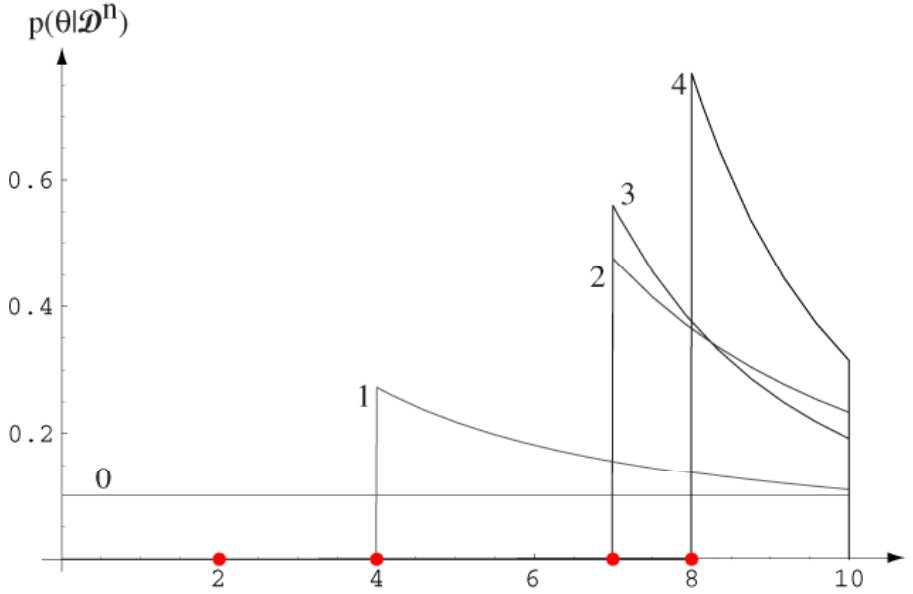
$$p(\theta \mid D^1) \propto p(x \mid \theta) p(\theta \mid D^0) = \begin{cases} 1/\theta & \gamma\iota\alpha \;\; 4 \le \theta \le 10 \\ 0 & \alpha\lambda\lambda o\acute{u} \end{cases}$$

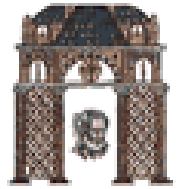
$$p(\theta \mid D^2) \propto p(x \mid \theta) p(\theta \mid D^1) = \begin{cases} 1/\theta^2 & \gamma\iota\alpha \;\; 7 \le \theta \le 10 \\ 0 & \alpha\lambda\lambda o\acute{u} \end{cases}$$

$$p(\theta \mid D^3) \propto p(x \mid \theta) p(\theta \mid D^2) = \begin{cases} 1/\theta^3 & \gamma\iota\alpha \;\; 7 \le \theta \le 10 \\ 0 & \alpha\lambda\lambda o\acute{u} \end{cases}$$

$$p(\theta \mid D^4) \propto p(x \mid \theta) p(\theta \mid D^3) = \begin{cases} 1/\theta^4 & \gamma\iota\alpha \;\; 8 \le \theta \le 10 \\ 0 & \alpha\lambda\lambda o\acute{u} \end{cases}$$

$$p(\theta \mid D^n) \propto p(x \mid \theta) p(\theta \mid D^{n-1}) = \begin{cases} 1/\theta^n & \gamma\iota\alpha \;\; \max[D^n] \le \theta \le 10 \\ 0 & \alpha\lambda\lambda o\upsilon \end{cases}$$

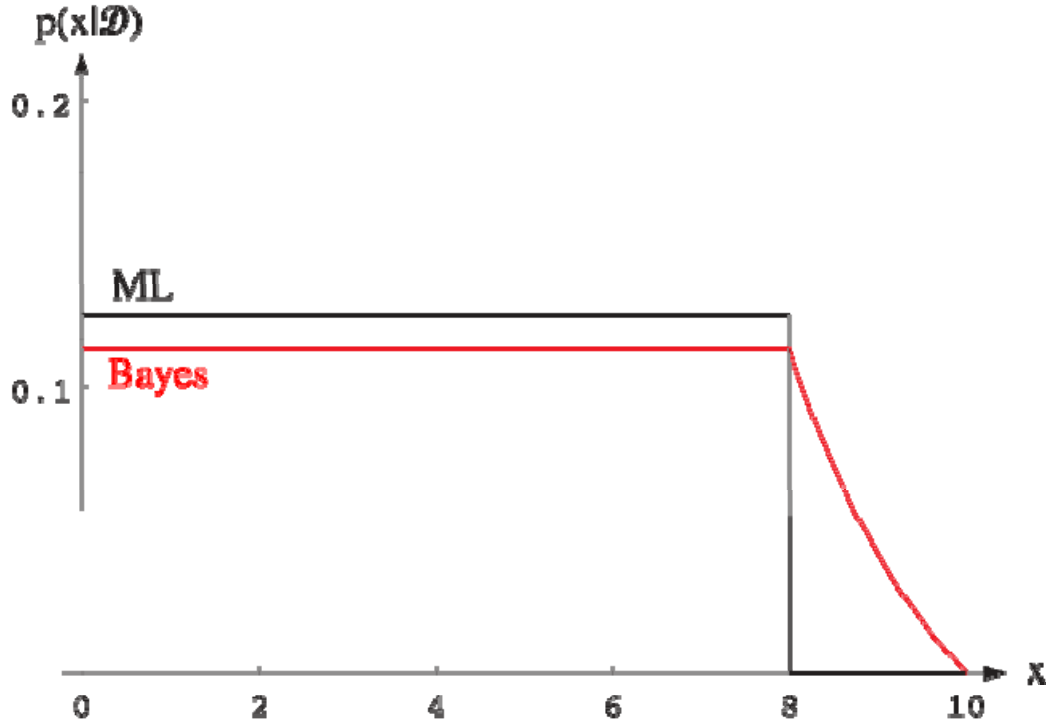




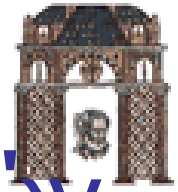
Παράδειγμα 1 (συν.): Αναδρομική Εκτίμηση Παραμέτρων

Άρα έχουμε με τα 4 σημεία ότι ο εκτιμητής του p(x|D) της μεγίστης πιθανοφάνειας είναι **η τιμή θ=8**, άρα p(x|D) = U(0,8),

ενώ ο εκτιμητής Bayes (κόκκινη γραμμή) του p(x|D) είναι $p(x \mid D) = \int p(x \mid \theta) p(\theta \mid D) d\theta$

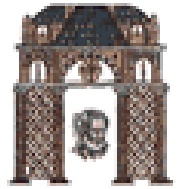


Διότι για 0≤x≤8, p(x|θ) ~1/8, για 8≤x≤10, p(x|θ) ~ 1/θ.



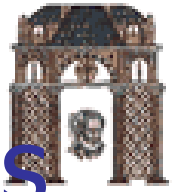
Διαφορές ML και Bayes εκτιμητών

- Υπολογιστική πολυπλοκότητα
- ML πιο κατανοητό αποτέλεσμα
- Bayes δεν οδηγεί απαραίτητα στο παραμετρικό μοντέλο που υποθέσαμε (δες τελευταίο παράδειγμα)
- Bayes χρησιμοποιεί καλύτερα τα δεδομένα εκπαίδευσης
- Bayes χρειάζεται και την εκ των προτέρων κατανομή (συνήθως δεν είναι γνωστή)



Curse of Dimentionality

Εάν για την καλή εκτίμηση των παραμέτρων μιας μονοδιάστατης pdf χρειάζονται N δείγματα, για την εκτίμηση της pdf σε d διαστάσεις χρειάζονται N^d δείγματα



Απλοποιημένος Εκτιμητής Bayes

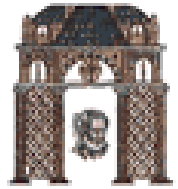
Στην περίπτωση αυτή υποθέτουμε ότι τα d χαρακτηριστικά ενός προτύπου $x=(x_1,x_2,\ldots,x_d)$ είναι στατιστικά ανεξάρτητα

Άρα

$$p(\mathbf{x}\,|\,\omega_i)=\prod_{n=1}^{d} p(x_n\,|\,\omega_i)$$

Η τεχνική αυτή εφαρμόζεται όταν το d είναι μεγάλο (curse of dimensionality) και ο αριθμός N των δεδομένων εκπαίδευσης είναι σχετικά μικρός.

Συνήθως υποθέτουμε ότι τα pdf των x_1, x_2, ..., x_d είναι κανονικές κατανομές και προσπαθούμε να εκτιμήσουμε την μέση τιμή και την διασπορά τους (μ_n,σ_n^2)



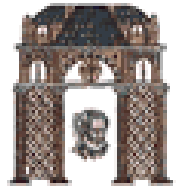
Παράδειγμα

Έστω δεδομένα που δημιουργούνται από την παρακάτω pdf

$$p(\mathbf{x} \mid \omega_i) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{d}{2}}} \exp\left(-(\mathbf{x}-\mu_\mathbf{i})^T \Sigma_i^{-1} (\mathbf{x}-\mu_\mathbf{i})\right)$$

Με i=2, d=5 και $\mu_1=[0\ 0\ 0\ 0\ 0]^T$, $\mu_2=[1\ 1\ 1\ 1\ 1]T$

$$\Sigma_1 = \begin{bmatrix} 0{,}8 & 0{,}2 & 0{,}1 & 0{,}05 & 0{,}01 \\ 0{,}2 & 0{,}7 & 0{,}1 & 0{,}03 & 0{,}02 \\ 0{,}1 & 0{,}1 & 0{,}8 & 0{,}02 & 0{,}01 \\ 0{,}05 & 0{,}03 & 0{,}02 & 0{,}9 & 0{,}01 \\ 0{,}01 & 0{,}02 & 0{,}01 & 0{,}01 & 0{,}8 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 0{,}9 & 0{,}1 & 0{,}05 & 0{,}02 & 0{,}01 \\ 0{,}1 & 0{,}8 & 0{,}1 & 0{,}02 & 0{,}02 \\ 0{,}05 & 0{,}1 & 0{,}7 & 0{,}02 & 0{,}01 \\ 0{,}02 & 0{,}02 & 0{,}02 & 0{,}6 & 0{,}02 \\ 0{,}01 & 0{,}02 & 0{,}01 & 0{,}02 & 0{,}7 \end{bmatrix}$$



MIXTURE MODELS

Μία τυχαία pdf μπορεί να παρασταθεί σαν άθροισμα γνωστών pdf.

$$p(x) = \sum_{j=1}^{J} P_j \, p(x \mid j) \quad \text{όπου}$$

$$\sum_{j=1}^{J} P_j = 1, \qquad \int_{x=-\infty}^{\infty} p(x|j)dx = 1$$

Μία συνήθης επιλογή για τις $p(x/j)$ είναι η κανονική $N(\mu_j, \Sigma_j)$

Άρα δεδομένου ενός δείγματος τιμών προσπαθούμε να εκτιμήσουμε τους συντελεστές Pj και τις παραμέτρους των $p(x/j)$

Ένας δημοφιλής αλγόριθμος που υπολογίζει τις παραμέτρους μέσα από μία επαναληπτική διαδικασία είναι ο ΕΜ (δες Θεοδωρίδη, 4-η έκδοση, κεφ, 2.5.5)