

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Μη-παραμετρικοί Ταξινομητές

K Πλησιέστερα Γειτονικά Στοιχεία (*K*-Nearest Neighbors - KNN)

Στατιστική Αξιολόγηση Δυναμικής Ταξινόμησης

Μετρικές Αξιολόγησης Μεθόδων, Μήτρα Σύγχυσης, ROC, AUC

Παραμετρική Πιθανοτική Ταξινόμηση

Εκτίμηση MLE & MAP, Ταξινομητής Bayes, Αλγόριθμος Naive Bayes

καθ. Βασίλης Μάγκλαρης

maglaris@netmode.ntua.gr

www.netmode.ntua.gr

Αίθουσα 002, Νέα Κτίρια ΣΗΜΜΥ

Τρίτη 16/5/2023

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

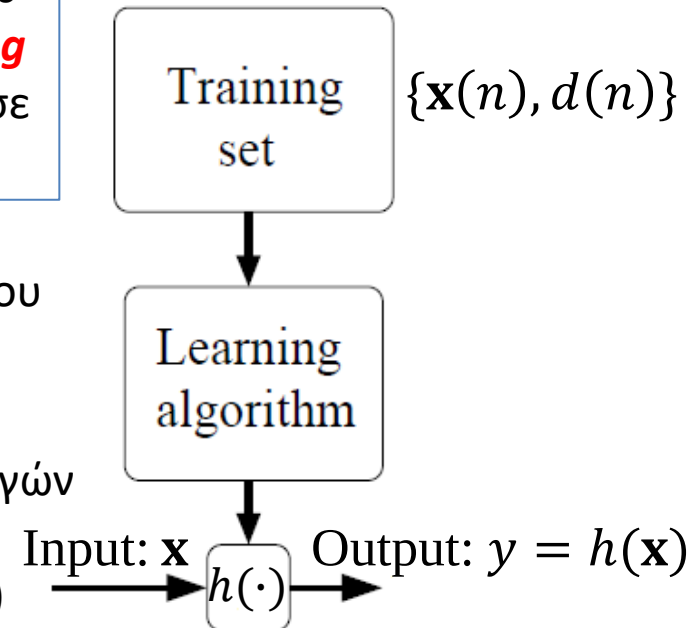
Γενικό Μοντέλο Επιβλεπόμενης Μάθησης - Supervised Learning (επανάληψη)

Βασισμένο στο Andrew Ng, "CS229 Lecture Notes", Stanford University, Fall 2018

- Στόχος του συστήματος είναι η αντιστοίχιση ενός δειγματικού στοιχείου εισόδου (**input sample point, example, instance**) $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_m]^T$ σε τιμές εξόδου y που εκτιμούν επιθυμητές τιμές d (**labels, targets**) π.χ. πρόβλεψη ή ταξινόμηση. Τα στοιχεία x_i είναι αριθμητικές τιμές που κωδικοποιούν m ειδοποιά χαρακτηριστικά (**features**) του δειγματικού στοιχείου $\mathbf{x}$

Ζητείται ο προσδιορισμός της συνάρτησης εισόδου - εξόδου $y = h(\mathbf{x}) \cong d$ που προκύπτει από δείγμα μάθησης (**Training Set**) N **labeled** ζευγών $\{\mathbf{x}(n), d(n)\}$, $n = 1, 2, \dots, N$ γνωστών σε εξωτερικό εκπαιδευτή (**supervisor**)

- Η μορφή και οι παράμετροι της $h(\cdot)$ προσδιορίζονται με αλγόριθμο μάθησης που συγκλίνει σε προσέγγιση του στόχου της υπόθεσης για τα N στοιχεία του δείγματος μάθησης $d(n) \cong y(n) = h(\mathbf{x}(n))$
- Αν ο στόχος ικανοποιείται με μικρό αριθμό διακριτών επιλογών (κλάσεων) της y πρόκειται για πρόβλημα Ταξινόμησης, **Classification** (για δύο κλάσεις έχουμε δυαδική ταξινόμηση)
- Αν η έξοδος y λαμβάνει συνεχείς τιμές, το πρόβλημα αναφέρεται σαν Παλινδρόμηση, **Regression**



Ταξινόμηση σύμφωνα με Πλειοψηφία K Πλησιέστερων Γειτονικών Στοιχείων (1/2)

Μη Παραμετρική Μέθοδος: Δεν θεωρεί πιθανοτικό μοντέλο δείγματος αλλά βασίζεται σε εκτιμήσεις *απόστασης* από στοιχεία μάθησης με γνωστά *labels*. Πλεονέκτημα όταν δεν είναι προσιτή η στατιστική δομή του δείγματος μάθησης, π.χ. ιστόγραμμα, αλλά συνήθως απαιτεί μεγαλύτερο αριθμό στοιχείων μάθησης από **Παραμετρικές Μεθόδους** εκτίμησης παραμέτρων μοντέλου πιθανοτήτων – π.χ. *Gauss*

Αλγόριθμος K -Nearest Neighbors (KNN)

Απαιτείται *labeled* δείγμα μάθησης $\{\mathbf{x}(n), d(n)\}$, $n = 1, 2, \dots, N$ από στοιχεία (πρότυπα) $\mathbf{x}(n) = [x_1(n) \ x_2(n) \ \dots \ x_m(n)]^T$, *labels* $d(n) \rightarrow \mathcal{C}$ της κλάσης του $\mathbf{x}(n)$, π.χ. $\mathcal{C} \in \{0, 1\}$ για δυαδική ταξινόμηση προτύπων, και μέτρο απόστασης π.χ. *Euclidean distance* $\|\mathbf{x}(i), \mathbf{x}(j)\| = \sqrt{\sum_{k=1}^m (x_k(i) - x_k(j))^2}$ ή απόσταση *Hamming* (ανόμοια δυαδικά ψηφία μεταξύ $\mathbf{x}(i), \mathbf{x}(j)$)

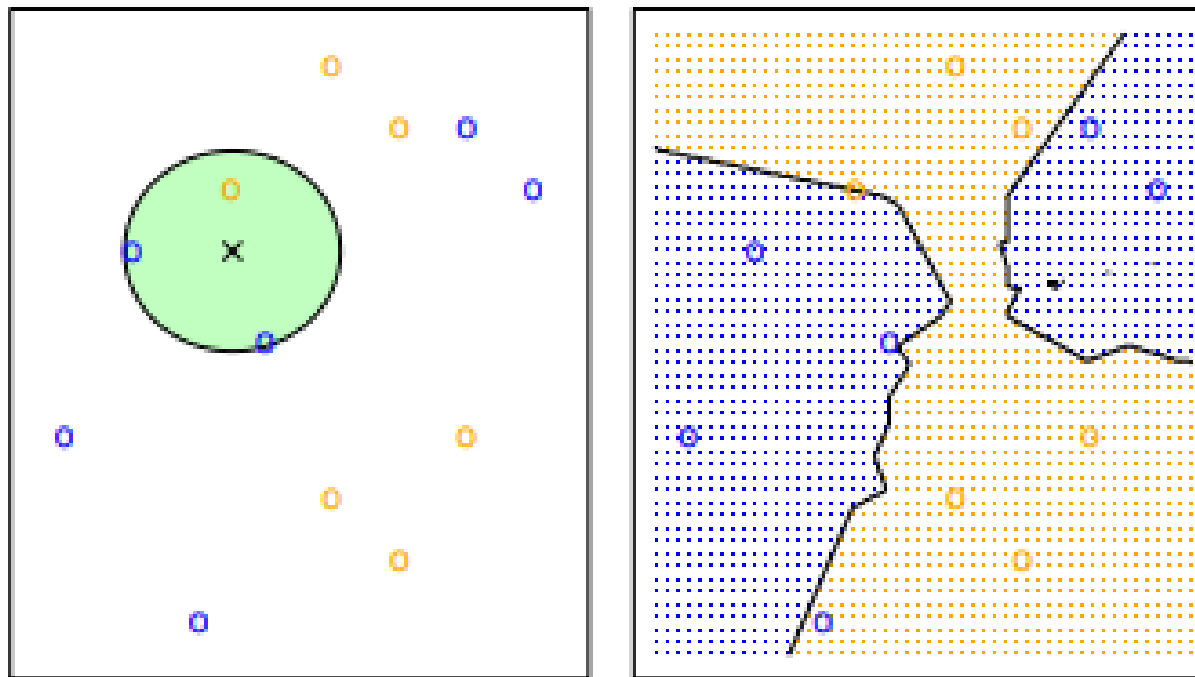
- **Φάση Μάθησης (*Lazy Learning Method*):** Αποθήκευση στη μνήμη του δείγματος μάθησης των N *labeled* διανυσμάτων - προτύπων $\{\mathbf{x}(n), d(n)\}$
- **Φάση Test:** Νέο στοιχείο $\hat{\mathbf{x}}$ ταξινομείται σύμφωνα με τη κλάση (*label*) της *πλειοψηφίας* των K πλησιέστερων διανυσμάτων από τα N στοιχεία μάθησης - πρότυπα $\mathbf{x}(n)$

K : Κρίσιμη **υπερπαραμέτρος**, φυσικός αριθμός (περιττός για δυαδική ταξινόμηση) που συνήθως απαιτεί δοκιμές ακρίβειας (π.χ. μέσω *cross validation*) για τον προσδιορισμό του

- $K = 1$: Νέο δειγματικό στοιχείο $\hat{\mathbf{x}}$ ταξινομείται σύμφωνα με τη κλάση του πλησιέστερου γείτονα του στο δείγμα μάθησης
- $K \gg 1$: Αντοχή σε παραμορφώσεις/θορύβους - εξομάλυνση περιοχών κατάταξης *αλλά* με αποθηκευτικό κόστος και πιθανή βλαβερή επιρροή *outliers* (στοιχείων με ασυνήθη χαρακτηριστικά) $\rightarrow$ κανονικοποίηση *dataset*, φιλτράρισμα συνιστωσών, διαγραφή *outliers*

Ταξινόμηση σύμφωνα με Πλειοψηφία K Πλησιέστερων Γειτονικών Στοιχείων (2/2)

https://hastie.su.domains/ISLR2/ISLRv2_website.pdf



Δυναδική Ταξινόμηση

$\mathcal{C} \in \{\text{Blue, Orange}\}$

- Blue Πρότυπα: ○
- Orange Πρότυπα: ○
- Σημείο Test: x

FIGURE 2.14. The KNN approach, using $K = 3$, is illustrated in a simple situation with six blue observations and six orange observations. Left: a test observation at which a predicted class label is desired is shown as a black cross. The three closest points to the test observation are identified, and it is predicted that the test observation belongs to the most commonly-occurring class, in this case blue. Right: The KNN decision boundary for this example is shown in black. The blue grid indicates the region in which a test observation will be assigned to the blue class, and the orange grid indicates the region in which it will be assigned to the orange class.

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Μετρικές Αξιολόγησης Στατιστικής Ταξινόμησης: Confusion Matrix, ROC, AUC (1/3)

Στατιστική Δυαδική Ταξινόμηση - Παραμετρικοί Ταξινομητές

Αντιστοίχιση παραδειγμάτων (στοιχείων) δείγματος σε 2 κλάσεις $\mathcal{C} \in \{P, N\}$:

Θετική (**Positive P**) - Αρνητική (**Negative N**)

- Διάγνωση μολύνσεων: **Θετικό** $\triangleq$ **Μολυσμένο Δειγματικό Στοιχείο**
- Διάγνωση (ανίχνευση) ανωμαλιών: **Θετικό** $\triangleq$ **Ανώμαλο Δειγματικό Στοιχείο**
- Αναγνώριση δυαδικών προτύπων (π.χ. γάτες – σκυλιά): **Θετικό** $\triangleq$ **Γάτα**, **Αρνητικό** $\triangleq$ **Σκύλος**

Οι **παραμετρικοί** αλγόριθμοι ταξινόμησης προκύπτουν από στατιστική εκτίμηση κατανομής του δείγματος μάθησης (π.χ. θεώρηση κατανομής **Gauss**, πρόβλεψη παραμέτρων από **labeled** δείγμα μάθησης σε **supervised learning** μέσω **regression** και σύγκριση με δυαδικό κατώφλι - **threshold** ή σιγμοειδή συνάρτηση - **logistic function**)

<https://www.section.io/engineering-education/parametric-vs-nonparametric/>

Παραμετρικές Μέθοδοι: Linear & Logistic Regression, Perceptron Convergence, Bayes Classifiers...

Μη-παραμετρικές Μέθοδοι: KNN, Decision Trees, SVM...

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Μετρικές Αξιολόγησης Στατιστικής Ταξινόμησης: Confusion Matrix, ROC, AUC (2/3)

Μήτρα Σύγχυσης - Confusion Matrix

- Λανθασμένες Προβλέψεις : False Positives - **FP**, False Negatives - **FN**
- Ορθές Προβλέψεις: True Positives - **TP**, True Negatives - **TN**

Ρυθμοί (Rates) Ορθών/Λανθασμένων Προβλέψεων :

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}} = 1 - \text{TPR}, \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} = 1 - \text{TNR}$$

		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Confusion Matrix

Μετρικές Αξιολόγησης Ταξινομητή

Accuracy (Ακρίβεια): $\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$ (λόγος συνολικά σωστών προβλέψεων)

Sensitivity, Recall (Ευαισθησία): $\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$ (σωστές προβλέψεις σε θετικά παραδείγματα)

Precision (Θετική Ακρίβεια): $\text{PRE} = \frac{\text{TP}}{\text{TP} + \text{FP}}$ (σωστές προβλέψεις από θετικές προβλέψεις)

F1-Score: $\text{F1} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})} = \frac{2}{\text{TPR}^{-1} + \text{ACC}^{-1}}$ (αρμονικός μέσος όρος ευαισθησίας & ακρίβειας)

Παράδειγμα Αναγνώρισης Εικόνας: Γάτα ή Σκύλος

https://en.wikipedia.org/wiki/Confusion_matrix

Test Sample 12 εικόνων: 8 γάτες (class P), 4 σκύλοι (class N)

Ο ταξινομητής μετά από στάδιο μάθησης προβλέπει 7 γάτες και 5 σκύλους (9 σωστά και 3 λάθη) όπως φαίνεται στη **Confusion Matrix**

Predicted class \ Actual class	Cat	Dog
Cat	6	2
Dog	1	3

- Ακρίβεια (Accuracy): $ACC = \frac{TP+TN}{TP+TN+FP+FN} = \frac{6+3}{12} = 3/4$
- Ευαισθησία (Sensitivity): $TPR = \frac{TP}{TP+FN} = \frac{6}{6+2} = 3/4$

Receiver Operating Characteristics (ROC), Area Under the Curve (AUC)

<https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>

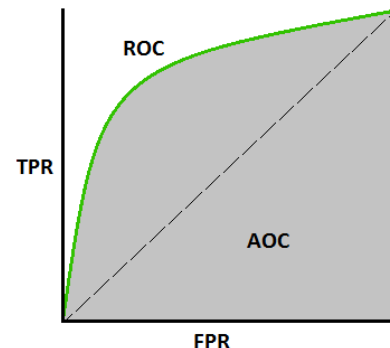
Ορισμοί από αναγνώριση στόχων σε δέκτες radar του Β' Παγκοσμίου Πολέμου
Λειτουργικές επιλογές διαχωρισμού (**threshold values**) σε σύστημα δυαδικής ταξινόμησης (**Positive** – Θετική, **Negative** – Αρνητική Πρόβλεψη) ανάλογα με τις προτιμήσεις του διαχειριστή μέσω σημείων **Receiver Operating Characteristics (ROC)**

- Διάγραμμα **ROC**: Συνάρτηση $FPR \rightarrow TPR$, $\{0 \leq FPR \leq 1, 0 \leq TPR \leq 1\}$
- Καλές λειτουργικές επιλογές **ROC**: $TPR \gg FPR$
- Ιδανικό σημείο: $TPR = 1, FPR = 0$

Μέτρο διαχωριστικής ικανότητας ταξινομητή

AUC: Εμβαδόν (επιφάνεια) της **ROC** για $0 \leq FPR \leq 1$

- Μη διαχωριστική ικανότητα: **AUC** = 0.5
- Διαχωριστική δεινότητα: **AUC** $\gg$ 0.5



Κανόνας **Bayes**

$$P(y|\mathbf{x}) = \frac{P(\mathbf{x}, y)}{P(\mathbf{x})} = \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})}$$

Παραμετρικό Μοντέλο Εκτίμησης Bayes

- Θεωρώ πως τα στοιχεία $\mathbf{x}(i) \in \{X\}$ είναι κατανεμημένα σύμφωνα με γνωστή κατανομή (π.χ. **Gauss**) σε **κλάσεις ταξινόμησης** $\mathcal{C}$ με βάση **διακριτές πιθανοτικές παραμέτρους** θ του δειγματικού χώρου, π.χ. μέση τιμή των $\mathbf{x}(i)$ για κάθε κλάση
- Οι εκτιμήσεις $\hat{\theta}$ των παραμέτρων θ υποδεικνύουν συμμετοχή σε **κλάσεις ταξινόμησης** παραδειγμάτων $\mathbf{x}(i)$
- Οι $\hat{\theta}$ προκύπτουν από παρατηρήσεις στοιχείων $\mathbf{x}(i)$ ενός υποσύνολου (**δείγμα μάθησης**) $\mathcal{D}$ του δειγματικού χώρου, με εξόδους ή **labels** $d(i) = \theta$ γνωστές στον εκπαιδευτή (**επιβλεπόμενη μάθηση**) και με βάση τον κανόνα του **Bayes**:

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$

όπου

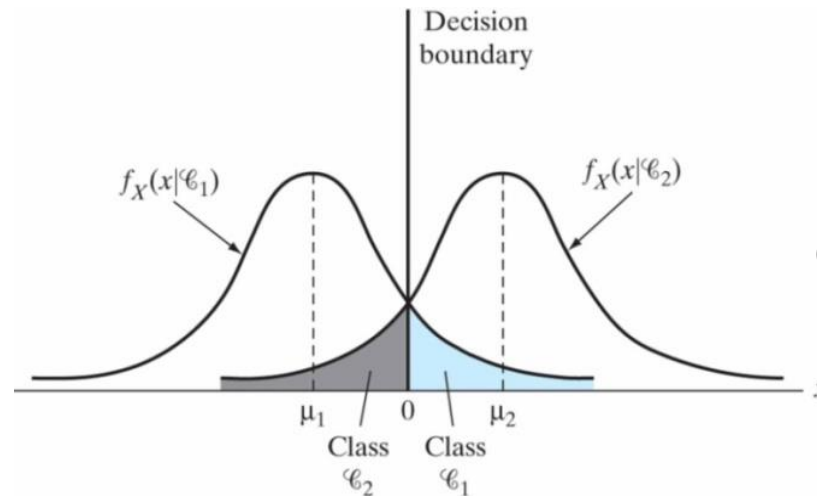
- $P(\mathcal{D})$: **Evidence**, πιθανότητα παραδειγμάτων $\mathbf{x}(i)$ υποσύνολου $\mathcal{D}$ δειγματικού χώρου $\{X\}$
- $P(\theta)$: **Prior**, πιθανότητα παραμέτρου θ δειγματικού χώρου $\{X\}$
- $P(\mathcal{D} | \theta)$: **Likelihood**, πιθανοφάνεια παραδειγμάτων στο $\mathcal{D}$ όταν η παράμετρος είναι θ
- $P(\theta | \mathcal{D})$: **Posterior**, ύστερη πιθανότητα παραμέτρου θ για παραδείγματα $\mathcal{D}$

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$

$P(\theta)$: *Prior Assumption* π.χ. εκτιμήσεις από εμπειρικές υποθέσεις για τον δειγματικό χώρο ή εκτιμήσεις από *labeled training set* $\mathcal{D}$

Παράδειγμα Παραμετρικού Δυναμικού Ταξινομητή

Υπόθεση: Κατανομή **Gauss** με μέση τιμή $\theta = \mu_1$ αν $x_n \rightarrow \mathcal{C}_1$ ή $\theta = \mu_2$ αν $x_n \rightarrow \mathcal{C}_2$



Likelihood Densities:

$$f_X(x|\mathcal{C}_1), f_X(x|\mathcal{C}_2)$$

Classification Errors:

Shaded areas

Δυο Κοινοί Τρόποι Εκτίμησης $\hat{\theta} \approx \theta$

1. Maximum Likelihood Estimation (**MLE**)

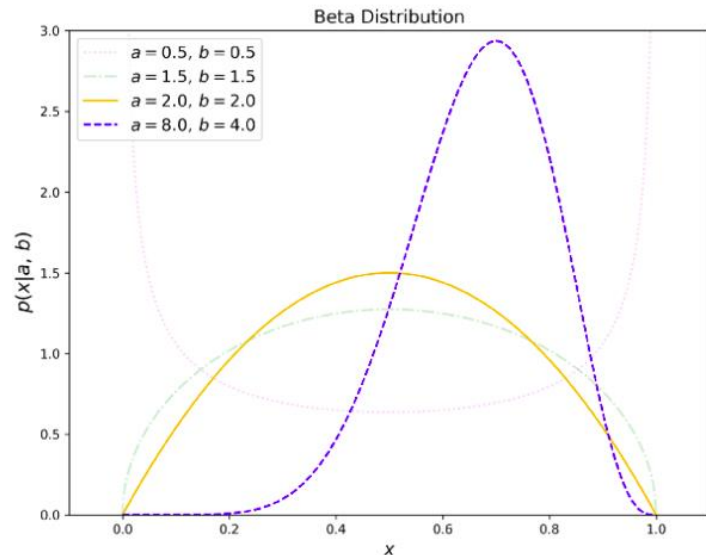
$$\hat{\theta} = \arg \max_{\theta} P(\mathcal{D}|\theta)$$

2. Maximum a Posteriori Probability (**MAP**) Estimation

$$\hat{\theta} = \arg \max_{\theta} P(\theta|\mathcal{D}) = \arg \max_{\theta} \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})} = \arg \max_{\theta} P(\mathcal{D}|\theta)P(\theta)$$

Παράδειγμα: Πείραμα Bernoulli για τυχαία μεταβλητή $X \in \{heads, tails\} \triangleq \{1,0\}$
 Δείγμα Μάθησης $\mathcal{D} = \{x(1), x(2), \dots, x(50)\}$ με 50 δοκιμές για εκτίμηση $\hat{\theta}$ της $\theta = P(X = 1)$, της πιθανότητας *heads*. Αν οι δοκιμές έβγαλαν $a_1 = 24$ *heads*, $a_0 = 26$ *tails*, η εκτίμηση **MLE** είναι $\hat{\theta} = \arg \max_{\theta} P(\mathcal{D}|\theta) = \frac{a_1}{(a_1+a_0)} = 0.48$, όπου $P(\mathcal{D}|\theta) = \theta^{a_1}(1-\theta)^{a_0}$

Για εκτίμηση **MAP** απαιτείται γνώση των *Prior* $P(\theta)$, π.χ. από εμπειρική παραδοχή για το δείγμα. Αν πιστεύουμε πως το νόμισμα είναι κάλπικο με $P(1) = 0.6$ μπορούμε να θεωρήσουμε $a_1 \rightarrow a_1 + \beta_1 = 24 + 9 = 33$, $a_0 \rightarrow a_0 + \beta_0 = 26 + 1 = 27$ και $\hat{\theta} = \frac{33}{33+27} = 0.55$



- Η *Likelihood* $P(\mathcal{D}|\theta) = \theta^{a_1}(1-\theta)^{a_0}$ (*Bernoulli*)
- Εκτιμούμε πως η *Prior* $P(\theta) \approx K\theta^{\beta_1-1}(1-\theta)^{\beta_0-1} = \text{Beta}(\beta_1, \beta_0) = \text{Beta}(9,1)$
- Θεωρούμε πως η *Posterior* $P(\theta|\mathcal{D})$ έχει την ίδια μορφή με την *Prior* $P(\theta)$ (**Conjugate Distributions**):

$$P(\theta) \sim \text{Beta}(\beta_1, \beta_0)$$

$$P(\theta|\mathcal{D}) \sim \text{Beta}(a_1 + \beta_1, a_0 + \beta_0)$$
- Η εκτίμηση $\hat{\theta} = \arg \max_{\theta} P(\theta|\mathcal{D})$ προκύπτει από τη μέγιστη τιμή της $\text{Beta}(33,27)$ για $\theta \cong 0.55$

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Πιθανοτικά Μοντέλα Ταξινόμησης, Παράδειγμα Ταξινομητή Bayes

<http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf>

Πιθανότητες $\sim$ Σχετική Συχνότητα Παραδειγμάτων $\{\mathbf{x}(i), d(i)\}$ στο Δείγμα Μάθησης $\mathcal{D}$

- Είσοδος $\mathbf{x}(i) = (Gender, HoursWorked)$ με 2 δυαδικές διαστάσεις (features)
- Έξοδος (label) $d(i) \cong y(i) = h(\mathbf{x}(i)) = Wealth$ δυαδική (poor, rich)

Gender	HoursWorked	Wealth	probability
female	< 40.5	poor	0.2531
female	< 40.5	rich	0.0246
female	≥ 40.5	poor	0.0422
female	≥ 40.5	rich	0.0116
male	< 40.5	poor	0.3313
male	< 40.5	rich	0.0972
male	≥ 40.5	poor	0.1341
male	≥ 40.5	rich	0.1059

Εκτιμήσεις Πιθανοτήτων
 $P(\mathbf{x}, y) = P(G, HW, y)$

όπου

$G \in \{M, F\}$

$HW \in \{light, hard\}$

$y \in \{poor, rich\}$

$$\text{Posterior } P(y|\mathbf{x}): P(rich|F, light) = \frac{0.0246}{0.2531+0.0246} \sim \mathbf{0.09}$$

Gender (G)	HrsWorked (HW)	$P(rich G, HW)$	$P(poor G, HW)$
F	<40.5 (<i>light</i>)	0.09	0.91
F	>40.5 (<i>hard</i>)	0.21	0.79
M	<40.5 (<i>light</i>)	0.23	0.77
M	>40.5 (<i>hard</i>)	0.38	0.62

$m = 2$ features $\{G, HW\}$ απαιτούν 4 εκτιμήσεις (m features απαιτούν 2^m εκτιμήσεις)

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Πιθανοτικά Μοντέλα Ταξινόμησης, Ταξινομητής Naïve Bayes (1/2)

<http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf>

Κανόνας του Bayes για Τυχαίες Μεταβλητές X, Y : $P(Y|X) = \frac{P(X,Y)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X)}$

Υπό Συνθήκη Ανεξαρτησία Τυχαίας Μεταβλητής $(X|Y, Z)$ από Y : $P(X|Y, Z) = P(X|Z)$

Προσεγγιστική Απλοποίηση – Naïve Bayes Classifier

Για δείγμα 2 χαρακτηριστικών $\mathbf{x} = [x_1 \ x_2]^T$ θεωρώ ότι τα x_i είναι ανεξάρτητα υπό τη συνθήκη της εξόδου y : $P(x_1|x_2, y) \cong P(x_1|y)$ οπότε:

$$P(\mathbf{x}|y) = P(x_1, x_2|y) = P(x_1|x_2, y) \times P(x_2|y) \cong P(x_1|y) \times P(x_2|y)$$

Γενικεύοντας την ανεξαρτησία υπό συνθήκη της εξόδου y για m χαρακτηριστικά (*features*) παραδείγματος $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_m]^T$ ισχύει για το *likelihood*

$$P(\mathbf{x}|y) = P(x_1, x_2, \dots, x_m | y) \cong \prod_{k=1}^m P(x_k | y)$$

Ο **Naïve Bayes Classifier** βασίζεται στην εκτίμηση της *posterior* $P(d|\mathbf{x}) \cong P(y|\mathbf{x})$ με βάση το *training sample*

$$P(y|\mathbf{x}) = \frac{P(y)P(\mathbf{x}|y)}{P(\mathbf{x})} \propto P(y)P(x_1|y)P(x_2|y) \dots P(x_m|y)$$

Απαιτούνται $\sim m$ εκτιμήσεις για ταξινόμηση ενός νέου παραδείγματος του δείγματος **test**: $\mathbf{x}^{new} = [x_1^{new} \ x_2^{new} \ \dots \ x_m^{new}]^T$ αντί για 2^m (αντιμετώπιση (;) του **curse of dimensionality**)

Οι εκτιμήσεις των *prior* $P(y)$ προκύπτουν από τη συχνότητα εμφάνισης στα παραδείγματα του δείγματος μάθησης (**Multinomial Naïve Bayes Classifier** για διακριτές τιμές των x_i) ή από παραδοχή Gauss (**Gaussian Naïve Bayes Classifier** για συνεχείς x_i)

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Πιθανοτικά Μοντέλα Ταξινόμησης, Ταξινομητής Naive Bayes (2/2)

<http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf>

Ο **Naive Bayes Classifier** βασίζεται στην προσέγγιση της **posterior** $P(d|\mathbf{x}) \cong P(y|\mathbf{x})$ σαν γινόμενο **ανεξαρτήτων** υπό συνθήκη **likelihoods** των χαρακτηριστικών (**features**)

$$P(y|\mathbf{x}) \propto P(y)P(x_1|y)P(x_2|y) \dots P(x_m|y)$$

Naive Bayes Algorithm:

Από το **labeled** δείγμα μάθησης $\mathcal{D} = \{(\mathbf{x}(1), d(1)), \dots, (\mathbf{x}(N), d(N))\}$ εκτιμώνται:

- Οι **prior** $P(d) \cong P(y)$ για όλες τις δυνατές κλάσεις d , π.χ. $d \in \{0,1\}$ για δυαδική ταξινόμηση
- Οι **likelihood** $P(x_k = l|y = d) \triangleq \theta_{kld}$ για κάθε (διακριτό) χαρακτηριστικό $k = 1, 2, \dots, m$ των στοιχείων μάθησης x_k που στο δείγμα μάθησης κατετάγη στη κλάση d

Νέο παράδειγμα του δείγματος **test** $\mathbf{x}^{new} = [x_1^{new} x_2^{new} \dots x_m^{new}]^T$, $x_k^{new} = l$ θα καταταγεί στη κλάση y^{new} που προκύπτει από τη σχέση:

$$y^{new} \leftarrow \arg \max_y P(y) \prod_{k=1}^m P(x_k^{new}|y)$$

ή

$$y^{new} \leftarrow \arg \max_d P(d) \prod_{k=1}^m \theta_{kld}$$

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Παράδειγμα Ταξινομητή Naive Bayes

<https://towardsdatascience.com/all-about-naive-bayes-8e13cef044cf>

Δείγμα Μάθησης - Labeled Sample από 1000 Στοιχεία (Training Examples)

Fruit	Long	Sweet	Yellow	Total
Banana	400 (80%)	350 (70%)	450 (90%)	500 (50%)
Orange	0 (0%)	150 (50%)	300 (100%)	300 (30%)
Other	100 (50%)	150 (75%)	50 (25%)	200 (20%)
Total	500 (50%)	650 (65%)	800 (80%)	1000

$$P(y|\mathbf{x}) \propto P(y)P(x_1|y)P(x_2|y) \dots P(x_m|y)$$

- $P(\text{Banana}|\text{Long, Sweet, Yellow}) \propto (0.5) \times (0.8) \times (0.7) \times (0.9) = 0.252$
- $P(\text{Orange}|\text{Long, Sweet, Yellow}) = 0$
- $P(\text{Other}|\text{Long, Sweet, Yellow}) \propto (0.2) \times (0.5) \times (0.75) \times (0.25) = 0.01875$

Ταξινόμηση Νέου Δειγματικού Στοιχείου (Test Example)

Φρούτα με χαρακτηριστικά $\mathbf{x} = (\text{Long, Sweet, Yellow})$ ανήκουν στην κλάση $y = (\text{Banana})$ με τη μεγαλύτερη posterior πιθανότητα $P(y|\mathbf{x}) \propto 0.252$

Σημείωση: Η τιμή της posterior μπορεί να εκτιμηθεί με κανονικοποίηση

$$P(y|\mathbf{x}) = \frac{0.252}{0.252 + 0.01875} = 0.931$$