ε.δε.μ²
ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ & ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

# ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

## Πιθανοτική Ταξινόμηση: Κανόνας Bayes & Αλγόριθμοι Naive Bayes


καθ. Βασίλης Μάγκλαρης

maglaris@netmode.ntua.gr

www.netmode.ntua.gr


Video Conference μέσω Cisco Webex

Πέμπτη 21/5/2020

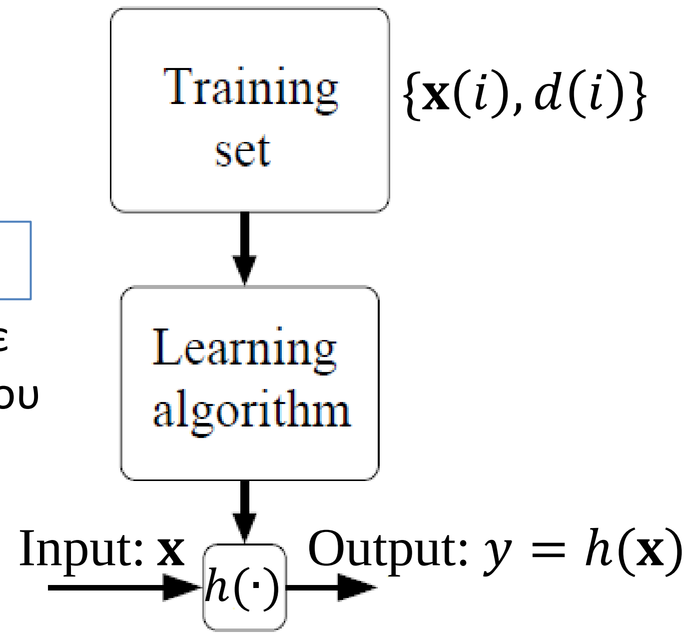
# ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

## Γενικό Μοντέλο Επιβλεπόμενης Μάθησης - Supervised Learning (*επανάληψη*)

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

- Στόχος του συστήματος είναι η αντιστοίχηση ενός δειγματικού στοιχείου εισόδου (***input sample point***, ***example***) $\mathbf{x} = [x_1\ x_2\ \dots x_m]^{\mathrm{T}}$ σε τιμές εξόδου $y$ που εκτιμούν επιθυμητή (***desired***) έξοδο $d$ συμβατή με δεδομένη υπόθεση (π.χ. πρόβλεψη ή ταξινόμηση). Τα στοιχεία $x_i$ είναι αριθμητικές τιμές που κωδικοποιούν $m$ ειδοποιά χαρακτηριστικά (***features***) του $\mathbf{x}$

Ζητείται ο προσδιορισμός της συνάρτησης $y = h(\mathbf{x}) \cong d$




- Η σχεδίαση της $h(\cdot)$ προκύπτει από αλγόριθμο μάθησης, με προσαρμογή της μορφής και των παραμέτρων ενός μοντέλου σε εμπειρικά δεδομένα ενός συνόλου $N$ χαρακτηρισμένων (***labeled***) του δείγματος μάθησης (***Training Set***) ζευγών, $\{\mathbf{x}(i), d(i)\}$, $i = 1,2, \dots, N$ ώστε να προσεγγίζεται ο στόχος της υπόθεσης $d(i) \cong y(i) = h(\mathbf{x}(i))$
- Αν ο στόχος ικανοποιείται με μικρό αριθμό διακριτών επιλογών της $y$ πρόκειται για πρόβλημα Ταξινόμησης, ***Classification*** (για δύο επιλογές έχουμε δυαδική ταξινόμηση)
- Αν η έξοδος $y$ λαμβάνει συνεχείς τιμές, το πρόβλημα αναφέρεται σαν Παλινδρόμηση, ***Regression***

# ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

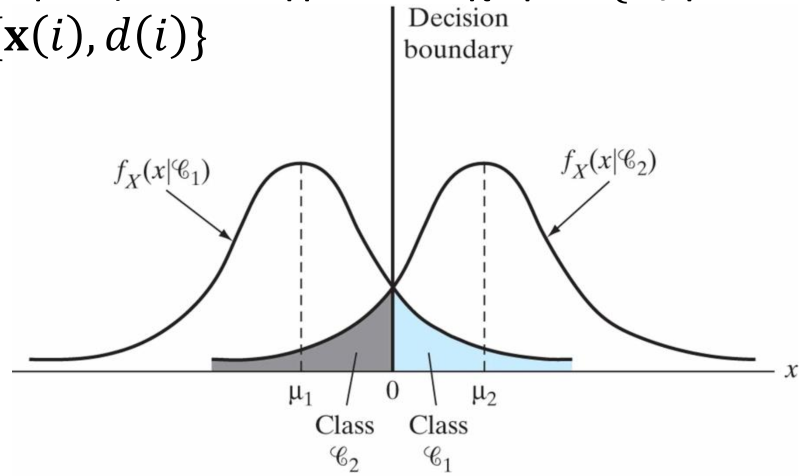
## Πιθανοτικά Μοντέλα Ταξινόμησης, Εκτίμηση Παραμέτρων MLE, MAP (1/2)

http://www.cs.cmu.edu/~tom/mlbook/Joint_MLE_MAP.pdf

$$P(y|\mathbf{x}) = \frac{P(\mathbf{x},y)}{P(\mathbf{x})} = \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})} \quad \text{(Κανόνας Bayes)}$$

$P(\mathbf{x})$: **Prior** (πρότερη) πιθανότητα παραδείγματος εισόδου $\mathbf{x}$ του δειγματικού χώρου $\{\mathbf{X}\}$
$P(y)$: **Class Prior** πιθανότητα εξόδου της τάξης $y$
$P(\mathbf{x}|y)$: **Likelihood** (πιθανοφάνεια) εισόδου $\mathbf{x}$ όταν η έξοδος υποδεικνύει την τάξη $y$
$P(y|\mathbf{x})$: **Posterior** (ύστερη) πιθανότητα ταξινόμησης στην τάξη $y$ παραδείγματος εισόδου $\mathbf{x}$

**Εκτίμηση παραμέτρων** θ (πιθανοτήτων, ροπών, κατανομών) του δειγματικού χώρου $\{\mathbf{X}\}$ με βάση τις παρατηρήσεις του δείγματος μάθησης $D = \{\mathbf{x}(i), d(i)\}$




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

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

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

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

$P(\theta)$: *Prior* Assumption

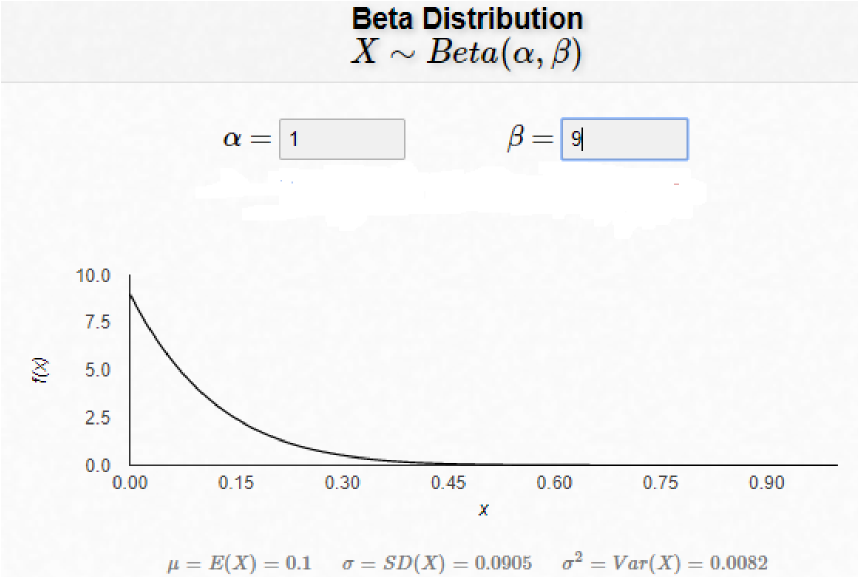
# ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

## Πιθανοτικά Μοντέλα Ταξινόμησης, Εκτίμηση Παραμέτρων MLE, MAP (2/2)

http://www.cs.cmu.edu/~tom/mlbook/Joint_MLE_MAP.pdf

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

**Παράδειγμα**: Πείραμα Bernoulli για τυχαία μεταβλητή $X = \{heads, tails\} \triangleq \{1{,}0\}$
Δείγμα Μάθησης $D = \{x(1), x(2), \ldots, x(50)\}$ 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 $\hat{\theta} = \frac{a_1}{(a_1+a_0)} = 0.48$. Για εκτίμηση **MAP** απαιτείται γνώση των $P(\theta)$, π.χ. από εμπειρική παραδοχή για το δείγμα. Αν πιστεύουμε πως το νόμισμα είναι κάλπικο με $P(1) = 0.6$ μπορούμε να θεωρήσουμε $a_1 \to 24+9$, $a_0 \to 26+1$ οπότε $\hat{\theta} \to 0.55$
Προκύπτει με *Prior* $P(\theta) = Beta(\beta_0, \beta_1) = K\theta^{\beta_1 - 1}(1-\theta)^{\beta_0 - 1} = Beta(1{,}9)$




Αν ξέρουμε πως οι όλες οι επιλογές της θ έχουν ίσες πιθανότητες, τότε **MAP** ≡ **MLE**

https://homepage.divms.uiowa.edu/~mbognar/applets/beta.html

# ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

## Πιθανοτικά Μοντέλα Ταξινόμησης, Κανόνας του Bayes (1/3)

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

Πιθανοτήτες ~ Σχετική Συχνότητα Παραδειγμάτων $\{\mathbf{x}(i), d(i)\}$ στο Δείγμα Μάθησης

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

| Gender | HoursWorked | Wealth | probability |
|---|---|---|---|
| female | $< 40.5$ | poor | 0.2531 |
| female | $< 40.5$ | rich | 0.0246 |
| female | $\geq 40.5$ | poor | 0.0422 |
| female | $\geq 40.5$ | rich | 0.0116 |
| male | $< 40.5$ | poor | 0.3313 |
| male | $< 40.5$ | rich | 0.0972 |
| male | $\geq 40.5$ | poor | 0.1341 |
| male | $\geq 40.5$ | rich | 0.1059 |

$$P(\mathbf{x}, y) = P(\text{G}, \text{HW}, y)$$
$$\text{G} = \{M, F\}$$
$$\text{HW} = \{light, hard\}$$
$$y = \{\text{poor}, \text{rich}\}$$

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

| Gender (G) | HrsWorked (HW) | *P*(rich\|G,HW) | *P*(poor\|G,HW) |
|---|---|---|---|
| *F* | <40.5 (*light*) | **0.09** | 0.91 |
| *F* | >40.5 (*hard*) | 0.21 | 0.79 |
| *M* | <40.5 (*light*) | 0.23 | 0.77 |
| *M* | >40.5 (*hard*) | 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, Κανόνας του Bayes (2/3)

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 – Naive Bayes Classifier

Οι τυχαίες μεταβλητές που κωδικοποιούν τα $m$ χαρακτηριστικά (*features*) παραδείγματος $\mathbf{x} = [x_1\ x_2\ \dots\ x_m]^{\mathrm{T}}$ υπό την συνθήκη εξόδου $y \approx d$ είναι υπό συνθήκη ανεξάρτητες οπότε για το *likelihood* ισχύει

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

Ο ***Naive 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 $\mathbf{x}^{new} = [x_1^{new}\ x_2^{new}\ \dots\ x_m^{new}]^{\mathrm{T}}$ αντί για $2^m$ (αντιμετώπιση (;) του ***curse of dimensionality***)

Οι εκτιμήσεις των πιθανοτήτων προκύπτουν από συχνότητα εμφάνισης στα παραδείγματα στο δείγμα μάθησης (***Multinomial Naive Bayes Classifier*** - συνήθως για **διακριτές** τιμές των $x_i$) ή από παραδοχή Gauss (***Gaussian Naive Bayes Classifier*** - συνήθως για **συνεχείς** $x_i$)

# ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

## Πιθανοτικά Μοντέλα Ταξινόμησης, Κανόνας του Bayes (3/3)

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*)

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

**Naive Bayes Algorithm**:
Από το labeled δείγμα μάθησης $D = \{\mathbf{x}(i), d(i)\}, i = 1,2, \ldots, N$ εκτιμώνται:

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

Νέο δειγματικό στοιχείο $\mathbf{x}^{new} = [x_1^{new}\ x_2^{new} \ldots x_m^{new}]^{\mathrm{T}}$, $x_k^{new} = l$ 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

$$y^{new} \leftarrow \arg\max_{c} \pi_c \prod_{k=1}^{m} \theta_{klc}$$

# ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

## Παράδειγμα του Naive Bayes

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

**Δείγμα Μάθησης - Labeled Sample από 1000 Στοιχεία (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** |

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

- $P(\text{Banana}|\text{Long, Sweet, Yellow}) = \frac{(0.8)\times(0.7)\times(0.9)\times(0.5)}{(0.25)\times(0.33)\times(0.41)} = 0.252$
- $P(\text{Orange}|\text{Long, Sweet, Yellow}) = 0$
- $P(\text{Other}|\text{Long, Sweet, Yellow}) = \frac{(0.5)\times(0.75)\times(0.25)\times(0.2)}{(0.25)\times(0.33)\times(0.41)} = 0.01875$

**Ταξινόμηση Νέου Δειγματικού Στοιχείου (Test Example)**

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