ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Ενότητα 9 – Supervised Learning
Decision Trees (Δέντρα αποφάσεων)
Part 1: Classification

Μέθοδοι Μηχανικής Μάθησης στα Χρηματοοικονομικά

Αθανάσιος Σάκκας, ΟΠΑ

Decision Trees (Δέντρα αποφάσεων)

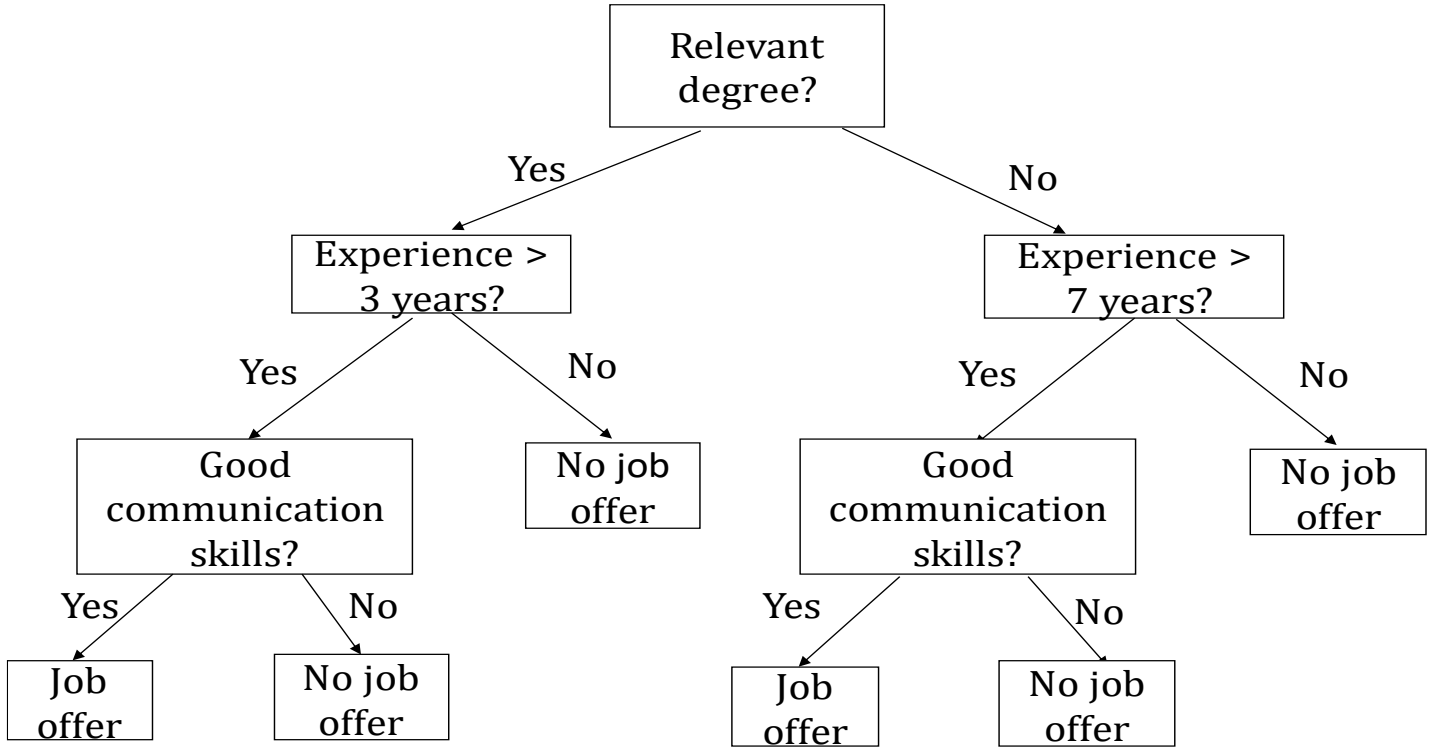
Πλεονεκτήματα σε σχέση με linear ή logistic regression

A. Ανταποκρίνονται στο πως σκέφτεται ο άνθρωπος και μπορούν εύκολα να εξηγηθούν και σε μη- εξειδικευμένους ανθρώπους.
B. Δεν απαιτείται η σχέση μεταξύ του target και των χαρακτηριστικών να είναι γραμμική
C. Το Decision Tree αυτομάτως επιλέγει τα καλύτερα χαρακτηριστικά για να κάνει πρόβλεψη.
D. Το Decision Tree είναι λιγότερο ευαίσθητο σε ακραίες τιμές σε σχέση με μία παλινδρόμηση.

- Στο πρώτο μέρος (Part 1), θα αναλύσουμε τη χρήση των Decision Trees για classification. Θα χρησιμοποιήσουμε την **Εφαρμογή LendingClub** που χρησιμοποιήσαμε και στην ενότητα με τη logistic regression.

- Ας ξεκινήσουμε με τη «φύση» των Decision Trees. Ας δούμε το παράδειγμα στην επόμενη διαφάνεια.

Παράδειγμα δέντρου αποφάσεων (Decision Tree) για τον καθορισμό του κριτηρίου για πρόσληψη



Το decision tree κοιτάει για ένα χαρακτηριστικό κάθε φορά και όχι όλα μαζί. Το πιο σημαντικό χαρακτηριστικό είναι να έχει σχετικό πτυχίο (relative degree). Μετά από αυτό έπονται τα χρόνια εμπειρίας και μετά οι επικοινωνιακές δεξιότητες.

Μέτρα αβεβαιότητας (Uncertainty)

- Υποθέστε ότι υπάρχουν n πιθανά αποτελέσματα και p_i η πιθανότητα του αποτελέσματος i όπου $\sum_{i=1}^{n} p_i = 1$

- **Entropy measure of uncertainty:** $-\sum_{i=1}^{n} p_i \log(p_i)$

  (Σημείωση: log είναι συνήθως log με βάση το 2 στη Μηχανική Μάθηση)

- **Gini Measure of uncertainty:** $\text{Gini} = 1 - \sum_{i=1}^{n} p_i^2$

  - Ο **δείκτης Gini** μπορεί να ερμηνευτεί ως το πόσο συχνά θα κάναμε λάθος στην ταξινόμηση μιας παρατήρησης, αν την κατατάσσαμε τυχαία σύμφωνα με την κατανομή των κατηγοριών στο σύνολο δεδομένων *ή πιο απλά*
  - Ο δείκτης Gini εκφράζει την πιθανότητα λανθασμένης ταξινόμησης μιας παρατήρησης, αν η κατηγορία της αποδοθεί με βάση την κατανομή των κατηγοριών στο δείγμα.

Κέρδος πληροφοριών (Information Gain)

- Το information gain είναι η αναμενόμενη μείωση στην αβεβαιότητα (entropy ή Gini).
- Υποθέστε ότι αρχικά υπάρχει 20% πιθανότητα κάποιος να έχει job offer (και 80% να μην έχει job offer.)
- Υποθέστε επιπλέον ότι υπάρχει 50% πιθανότητα ο ίδιος να έχει σχετικό πτυχίο (degree). Αν έχει σχετικό πτυχίο η πιθανότητα να έχει job offer αυξάνεται στο 30%, διαφορετικά πέφτει στο 10%.
- Initial entropy $= -[0.2\log(0.2) + 0.8\log(0.8)]=0.7219$

  Expected entropy $= -0.5[0.1\log(0.1) + 0.9\log(0.9)] - 0.5[0.3\log(0.3) + 0.7\log(0.7)]=0.6751$

  Expected information gain γνωρίζοντας ότι υπάρχει σχετικό πτυχίο $= 0.7219 - 0.6751=0.0468$
- Initial Gini $= 1-0.2^2-0.8^2=0.32$

  Expected Gini $= 0.5(1-0.1^2-0.9^2)+0.5(1-0.3^2-0.7^2)=0.30$

  Expected information gain γνωρίζοντας ότι υπάρχει σχετικό πτυχίο $= 0.32 - 0.30=0.02$

Decision Tree Algorithm

- Ο αλγόριθμος επιλέγει το χαρακτηριστικό (feature) στη ρίζα του δέντρου (tree) που έχει το μεγαλύτερο αναμενόμενο information gain.
- Παρομοίως, ψάχνει για το χαρακτηριστικό (feature) με το μεγαλύτερο information gain στους επόμενους κόμβους (nodes).
- Όταν υπάρχει όριο (threshold), καθορίζει το βέλτιστο όριο για κάθε χαρακτηριστικό (feature), δηλαδή, το όριο που μεγιστοποιεί το αναμενόμενο information gain για αυτό το χαρακτηριστικό, και βασίζει τους υπολογισμούς σε αυτό το όριο.

Εφαρμογή LendingClub

- Τα δεδομένα αποτελούνται από δάνεια, καλά (good) ή αθετημένα (defaulted).
- Εμείς θα χρησιμοποιήσουμε μόνο τέσσερα χαρακτηριστικά

1. Ιδιοκτησία σπιτιού (ενοικίαση έναντι ιδιοκτησία) - Home ownership (rent vs. own)
2. Εισόδημα – Income
3. Χρέος προς εισόδημα - Debt to income
4. Πιστωτική βαθμολογία - Credit score

- Το σύνολο δεδομένων έχει 12.290 παρατηρήσεις (9.733 καλά δάνεια “Fully Paid” και 2.557 δάνεια αθέτησης “Charged Off”). 7.000 δάνεια τέθηκαν σε training set, 3.000 σε validation set, και 2.290 σε test set.

Α. Επιλογή του ριζικού κόμβου όταν υπάρχουν τέσσερα χαρακτηριστικά

Στο training set υπάρχουν 5,542 καλά δάνεια (“Fully Paid”) και 1,458 δάνεια αθέτησης (“Charged Off”). Χωρίς καμία επιπλεόν πληροφορία η πιθανότητα για ένα καλό δάνειο εκτιμάται απο το training set ως 5,542/7000 = 79.17% και η πιθανότητα για ένα κακό δάνειο εκτιμάται ως 1,458 /7000 = 20.83% . Άρα η

Initial entropy = −0.7917×log(0.7917)−0.2083×log(0.2083) = 0.7382

A.1 Το πρώτο βήμα είναι να κατασκευάσουμε ένα δέντρο για να υπολογίσουμε το Information gain. Στο training set, 60.40% των δανειοληπτών έχουν δικό τους σπίτι και 39.60% ενοικιάζουν. Τα δάνεια που ήταν “Fully Paid” αφορά το 81.72% αυτών που είχαν δικό τους σπίτι και 75.29% αυτών που ενοικίαζαν.

Η **expected entropy** αν το home ownership (και κανένα άλλο χαρακτηριστικό) γίνεται γνωστό είναι

0.6040[-0.8172log(0.8172)-0.1828log(0.1828)]+0.3960[-0.7529log(0.7529)-0.2471log(0.2471)]=0.7339

Το information gain είναι 0.7382– 0.7339= 0.043

Στη συνέχεια υπολογίζουμε την expected entropy για το income. Στην περίπτωση αυτή χρειαζόμαστε κάποιο threshold income.

P_1 : Πιθανότητα το income να είναι πάνω από το threshold.

P_2 : Πιθανότητα αν το income είναι πάνω από το threshold, ο δανειολήπτης δε θα αθετήσει τις υποχρεώσεις του.

P_3: Πιθανότητα αν το income είναι κάτω από το threshold, ο δανειολήπτης θα αθετήσει τις υποχρεώσεις του.

Expected entropy = $P_1[- P_2 \log(P_2)-(1- P_2)\log(1- P_2)] +(1-P_1) [- P_3 \log(P_3)-(1- P_3)\log(1- P_3)]$.

Iterative search: Threshold income = \$48,079. P_1 = 70.84%, P_2 =81.15%, P_3 = 74.38%

Τα αποτελέσματα για τα optimal threshold values και το information again συνοψίζονται στον παρακάτω πίνακα

Feature	**Threshold value**	**Expected entropy**	**Expected Information gain**
Home Ownership	N.A.	0.7339	0.0043
Income ($'000s)	48.079	0.7342	0.0040
Debt to income ratio	19.85	0.7254	0.0128
FICO credit score	717.5	0.7270	0.0112

Το Debt to Income ratio με threshold=19.85 έχει το μεγαλύτερο information gain, άρα το Debt to Income ratio τοποθετείται στη ρίζα του δέντρου. Άρα τα αρχικά branches είναι Debt to Income ratio >19.85 και Debt to Income ratio <=19.85

Α2. Επαναλαμβάνουμε την διαδικασία για το επόμενο επίπεδο του δέντρου. Ο παρακάτω πίνακας δείχνει τα αποτελέσματα* για

Debt to Income ratio >19.85

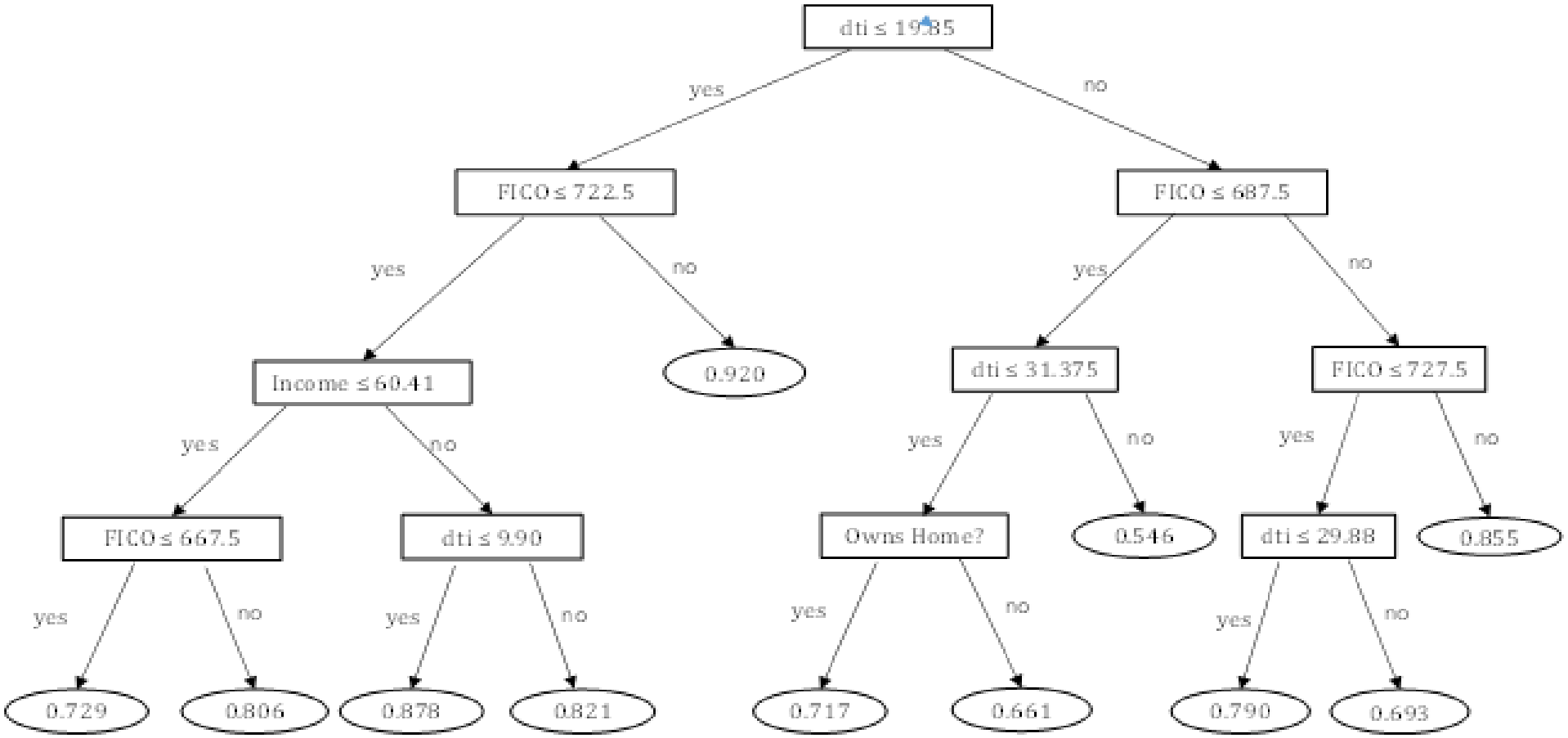
Feature	Threshold value	Expected entropy	Expected Information gain
Home Ownership	N.A.	0.8438	0.0060
Income ($'000s)	30.84	0.8463	0.0035
Debt to income ratio	33.87	0.8425	0.0072
FICO credit score	687.5	0.8361	0.0136

Debt to Income ratio ≤ 19.85

Feature	Threshold value	Expected entropy	Expected Information gain
Home Ownership	N.A.	0.6411	0.0030
Income ($'000s)	60.81	0.6409	0.0033
Debt to income ratio	10.74	0.6418	0.0023
FICO credit score	722.5	0.6340	0.0101

*Οι υπολογισμοί βρίσκονται στο excel *9.lending_club_Excel_decision_tree.xlsx*

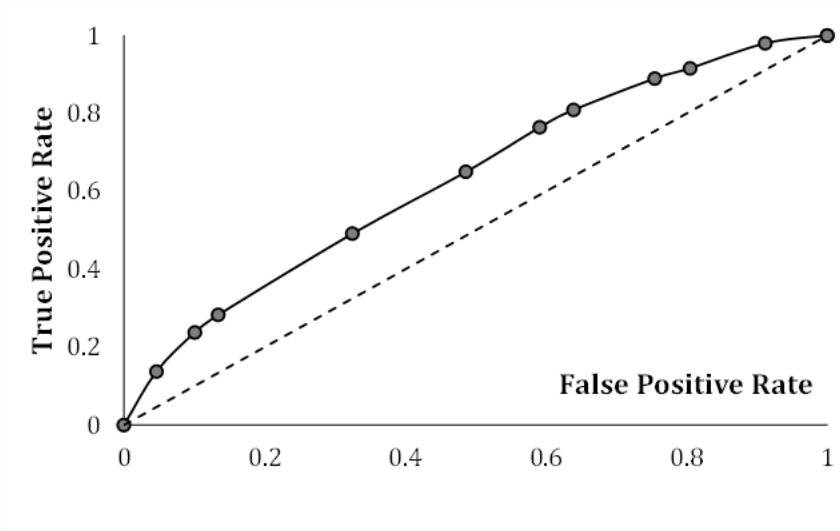
The Tree



Τα οβάλ σχήματα είναι τα leaves, δηλ. τα τελικά σημεία του δέντρου και δείχνουν το probability of default ή no default. Π.χ. Εάν το **Debt to Income ratio <= 19.85, και το FICO>722.5** το 0.920 (=739/803) είναι η estimated probability of no default.

Επιλογή κριτηρίου αποδοχής δανείων

- Όπως και με την logistic regression μπορούμε να επιλέξουμε να δεχτούμε δάνεια όπου η πιθανότητα ενός καλού δανείου είναι πάνω από κάποιο όριο (threshold) Z. Δείτε την ανάλυση στο *9.lending_club_Excel_decision_tree.xlsx*



AUC=0.6322

Python file: *9. lending_club_DecisionTree_Python.ipynb*

Στην python για την κατασκευή των δέντρων αποφάσεων χρησιμοποιούμε το sklearn.tree.DecisionTreeClassifier. Το documentation είναι το παρακάτω

https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Όπως θα δείτε και στο documentation για τον καθορισμού του decision tree. Θα χρειαστείτε να καθορίσετε τις τιμές των hyperparameters.

1. Στην περίπτωση του παραδείγματός μας θέσαμε το max depth = 4. Αυτό σημαίνει ότι θα προκύψουν το μέγιστο 4 επίπεδα στα οποία θα διαχωριστεί το δέντρο.
2. Ο ελάχιστος αριθμός παρατηρήσεων απαραίτητος για τον διαχωρισμό του δέντρου είναι 1,000 στο παραδειγμά μας: in_samples_split=1000. Η συγκεκριμένη hyperparameter εξηγεί γιατί το δέντρο μερικές φορές σταματάει πριν φτάσει το 4ο επίπεδο. Π.χ. Υπάρχουν μόλις 803 παρατηρήσεις όταν το dti<=19.85 και FICO>722.5. Παρομοίως υπάρχουν μόλις 304 παρατηρήσεις όταν dti>19.85 και FICO>722.5, και υπάρχουν μόλις 251 παρατηρήσεις όταν dti>31.375 και FICO<=687.5.

```
clf = DecisionTreeClassifier(criterion='entropy',max_depth=4,min_samples_split=1000,min_samples_leaf=200,random_state=0)
```

Bayes Theorem (χρήσιμο όταν θέλουμε μια εκτίμηση αβεβαιότητας καθώς και απλώς μια πρόβλεψη)

$$P(Y|X) = \frac{P(X\&Y)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X)}$$

Παράδειγμα: Παρατηρούμε ότι το 90% των δόλιων συναλλαγών (fraud) είναι για μεγάλα ποσά αργά την ημέρα (large & late). Επίσης, το 3% των συναλλαγών είναι για μεγάλα ποσά αργά την ημέρα (large & late) και το 1% των συναλλαγών είναι δόλιες (fraud).

$$P(\text{fraud}|\text{large\&late}) = \frac{P(\text{large\&late}|\text{fraud})P(\text{fraud})}{P(\text{large\&late})} = \frac{0.9 \times 0.01}{0.03} = 0.3$$

Bayes can be counterintuitive

- Ένα άτομο στις δέκα χιλιάδες έχει μια συγκεκριμένη ασθένεια.
- Μια εξέταση είναι 99% ακριβής (δηλαδή, εάν το άτομο έχει τη νόσο, το τεστ είναι στο 99% των φορών· ομοίως όταν το άτομο δεν έχει τη νόσο, το τεστ είναι σωστό το 99% των περιπτώσεων).
- Βρίσκεσαι θετικός.
- Ποια είναι η πιθανότητα να έχετε την ασθένεια;
- X=test positive, Y=has disease, $\bar{Y}$= does not have disease
- $P(X|Y) = 0.99;\ P(Y) = 0.0001$
- $P(X) = P(X|Y)P(Y) + P(X|\bar{Y})P(\bar{Y}) = 0.99 \times 0.0001 + 0.01 \times 0.9999 = 0.0101$
- $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{0.99 \times 0.0001}{0.0101} = 0.0098$

Naïve Bayes Classifier

Από το Bayes theorem

$$\text{Prob}(C|x_1, x_2, \cdots x_n) = \frac{\text{Prob}(x_1, x_2, \cdots x_n|C)}{\text{Prob}(x_1, x_2, \cdots x_n)} \text{Prob}(C)$$

If the features $\boldsymbol{x_i}$ are (approximately) independent this reduces to

$$\text{Prob}(C|x_1, x_2, \cdots x_n) = \frac{\text{Prob}(x_1|C)\text{Prob}(x_2|C) \cdots \text{Prob}(x_n|C)}{\text{Prob}(x_1, x_2, \cdots x_n)} \text{Prob}(C)$$

Παράδειγμα 1 (1)

- Η unconditional πιθανότητα ενός καλού δανείου είναι 0.85. Υπάρχουν 3 ανεξάρτητα (independent) χαρακτηριστικά (features).
- **Εάν ο αιτών έχει σπίτι (σημειώνεται με Η).** Η πιθανότητα ο αιτών να έχει δικό του σπίτι εάν το δάνειο είναι καλό είναι 60%, ενώ η πιθανότητα ο αιτών να έχει δικό του σπίτι εάν το δάνειο αθετηθεί είναι 50%.
- **Εάν ο αιτών εργάζεται για περισσότερο από ένα έτος (σημειώνεται με Ε).** Η πιθανότητα να απασχοληθεί ο αιτών για περισσότερο από ένα χρόνο εάν το δάνειο είναι καλό είναι 70% ενώ η πιθανότητα αυτό εάν το δάνειο αθετηθεί είναι 60%.
- **Εάν υπάρχουν δύο αιτούντες και όχι μόνο ένας (σημειώνεται με Τ).** Η πιθανότητα δύο αιτούντων όταν το δάνειο είναι καλό είναι 20% ενώ η πιθανότητα δύο αιτούντων όταν το δάνειο αθετηθεί είναι 10%.

Παράδειγμα 1 (2)

- $\text{Prob}(\text{Good Loan}|\text{H, E, T}) = \frac{0.6\times0.7\times0.2}{\text{Prob(H and E and T)}} \times 0.85 = \frac{0.0714}{\text{Prob(H and E and T)}}$

- $\text{Prob}(\text{Defaulting Loan}|\text{H, E, T}) = \frac{0.5\times0.6\times0.1}{\text{Prob(H and E and T)}} \times 0.15 = \frac{0.0045}{\text{Prob(H and E and T)}}$

- Αλλά αυτές οι πιθανότητες Prob(Good Loan) και Prob(Defaulting Loan) πρέπει να αθροίζονται στη μονάδα, δε χρειάζεται να υπολογίσουμε το Prob(H and E and T).
- Η πιθανότητα για ένα καλό δάνειο conditional στα H, E, and T είναι

$$\frac{0.0714}{0.0714 + 0.0045} = 0.941$$

- Η πιθανότητα για ένα αθετημένο δάνειο conditional στα H, E, and T είναι

$$\frac{0.0045}{0.0714 + 0.0045} = 0.059$$

- Το δέντρο στη διαφάνεια 12 μπορεί να θεωρηθεί και σαν ένα παράδειγμα Bayesian learning.

Παράδειγμα 2 (1)

Μπορούμε να χρησιμοποιήσουμε τον **Naïve Bayes Classifier** για συνεχείς κατανομές. Έστω ότι χρησιμοποιούμε τα δεδομένα από το LendingClub για να κάνουμε μια πρόβλεψη δανείου μέσω FICO και dti. Θεωρούμε ότι αυτά τα δύο χαρακτηριστικά είναι ανεξάρτητα (independent). Ο παρακάτω πίνακας δείχνει τη μέση τιμή και τυπική απόκλιση των FICO και dti *conditional* σε καλό δάνειο (good loan) και σε δάνειο που αθετήθηκε (default loan).

Loan result	**Mean FICO**	**SD FICO**	**Mean dti**	**SD dti**
Good loan	697.38	32.85	17.37	8.72
Defaulting loan	686.73	24.26	20.41	9.11

Υποθέστε κανονικές κατανομές και σκεφτείτε κάποιον που έχει βαθμολογία FICO = 720 και dti = 25.

Παράδειγμα 2 (2)

- Probability density for FICO conditional on good loan

$$\frac{1}{\sqrt{2\pi} \times 32.85} \exp\left(-\frac{(720 - 697.38)^2}{2 \times 32.85^2}\right) = 0.00958$$

- Probability density for dti conditional on good loan

$$\frac{1}{\sqrt{2\pi} \times 8.72} \exp\left(-\frac{(25 - 17.37)^2}{2 \times 8.72^2}\right) = 0.0312$$

- Probability density for FICO conditional default

$$\frac{1}{\sqrt{2\pi} \times 24.26} \exp\left(-\frac{(720 - 686.73)^2}{2 \times 24.26^2}\right) = 0.00642$$

- Probability density for dti conditional on default

$$\frac{1}{\sqrt{2\pi} \times 9.11} \exp\left(-\frac{(25 - 20.41)^2}{2 \times 9.11^2}\right) = 0.0385$$

Παράδειγμα 2 (3)

- Unconditional probability of a good loan είναι 0.8276
- Probability of a good loan conditional on FICO and dti είναι

$$\frac{0.00958 \times 0.0312 \times 0.7917}{Q} = \frac{2.366 \times 10^{-4}}{Q}$$

όπου Q is the probability density of FICO and dti.

- The probability of a defaulting loan conditional on FICO and dti είναι $\frac{0.00642 \times 0.0385 \times 0.2083}{Q} = \frac{0.516 \times 10^{-4}}{Q}$
- Probability of a good loan είναι $2.366/(2.366 + 0.516) = 0.821$.