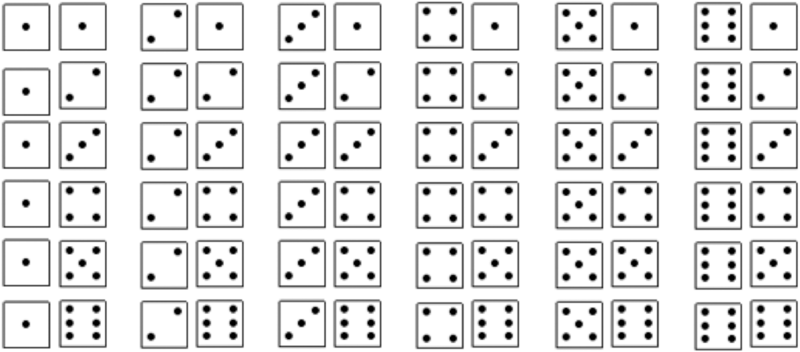
Παράρτημα Β: Εισαγωγή στη Θεωρία Πιθανοτήτων

Fundamentals of Machine Learning for Predictive Data Analytics

Αθανάσιος Σάκκας, ΟΠΑ

- Η θεωρία πιθανοτήτων είναι ο κλάδος των μαθηματικών που ασχολείται με τη μέτρηση της πιθανότητας (ή της αβεβαιότητας) γύρω από γεγονότα.
- Υπάρχουν δύο τρόποι να υπολογίσουμε την πιθανότητα ενός μελλοντικού γεγονότος:
  1. να χρησιμοποιήσουμε τη **σχετική συχνότητα** του γεγονότος στο παρελθόν,
  2. να χρησιμοποιήσουμε μια **υποκειμενική** εκτίμηση (ιδανικά από έναν ειδικό!).
- Θα εστιάσουμε στη χρήση της σχετικής συχνότητας.



Σχήμα: Ο **δειγματικός χώρος** (sample space) για το πεδίο δύο ζαριών.

ID	Die1	Die2
1	3	4
2	1	5
3	6	5
4	3	3
5	1	1

Πίνακας: Ένα dataset από τον δειγματικό χώρο του Σχήματος 1 [4].

- Σε όλη τη συζήτησή μας για πιθανότητες, θα μιλάμε για **γεγονότα** (events), οπότε είναι πολύ σημαντικό να κατανοείς αυτή την έννοια.

Γεγονότα (Events)

- Ένα **γεγονός** ορίζει μια ανάθεση τιμών στα features του πεδίου (domain). Αυτές οι αναθέσεις μπορεί να ορίζουν τιμές για όλα τα features (π.χ. μια πλήρη γραμμή στο dataset) ή μόνο για ένα/μερικά features.

Example

- Die1=*'3'*,
- Die1=*'1'*, Die2=*'5'*.

Συναρτήσεις Πιθανότητας: $P()$

- Ένα feature μπορεί να παίρνει μία ή περισσότερες τιμές από το πεδίο τιμών του και μπορούμε να βρούμε την πιθανότητα να πάρει μια συγκεκριμένη τιμή χρησιμοποιώντας μια **συνάρτηση πιθανότητας** $P()$.
- Μια συνάρτηση πιθανότητας είναι μια συνάρτηση που παίρνει ως παράμετρο ένα γεγονός (μια ανάθεση τιμών σε features) και επιστρέφει την πιθανότητα αυτού του γεγονότος.

Example

- $P(\text{Die1} = '3')$ επιστρέφει την πιθανότητα του γεγονότος $\text{Die1} = '3'$
- $P(\text{Die1} = '3', \text{Die2} = '4')$ επιστρέφει την πιθανότητα του γεγονότος όπου Die1='3' και Die2='4'.

Ιδιότητες Συναρτήσεων Πιθανότητας

$$\begin{aligned} 0 \leq P(f = level) &\leq 1 \\ \sum_i P(f = level_i) &= 1.0 \end{aligned}$$

- Οι συναρτήσεις πιθανότητας είναι πολύ εύκολο να δημιουργηθούν όταν έχουμε dataset.
- Η τιμή που επιστρέφει μια συνάρτηση πιθανότητας για ένα γεγονός είναι απλώς η **σχετική συχνότητα** του γεγονότος στο dataset — δηλαδή, πόσο συχνά συνέβη το γεγονός, διαιρεμένο με το πόσες φορές θα μπορούσε να συμβεί.

Example

- Η σχετική συχνότητα του γεγονότος Die1 $= '3'$ είναι το πλήθος των γραμμών όπου το feature παίρνει την τιμή 3, διαιρεμένο με το πλήθος των γραμμών του dataset.

Προγενέστερη Πιθανότητα (Prior / Unconditional)

- Η πιθανότητα ενός γεγονότος χωρίς καμία επιπλέον (συμφραζόμενη) πληροφορία.
- Το πλήθος των γραμμών όπου τα feature(s) παίρνουν τις σχετικές τιμές, διαιρεμένο με το συνολικό πλήθος γραμμών.

Example

$$P(\text{Die1} = '3') = \frac{|\{\mathbf{d}_1, \mathbf{d}_4\}|}{|\{\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, \mathbf{d}_4, \mathbf{d}_5\}|} = \frac{2}{5} = 0.4$$

Κοινή Πιθανότητα (Joint Probability)

- Η πιθανότητα δύο ή περισσότερων γεγονότων να συμβούν μαζί.
- Το πλήθος των γραμμών όπου ισχύει όλο το σύνολο αναθέσεων του joint event, διαιρεμένο με το συνολικό πλήθος γραμμών.

Example

$$P(\text{Die1} = \text{'6'}, \text{Die2} = \text{'5'}) = \frac{|\{\mathbf{d}_3\}|}{|\{\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, \mathbf{d}_4, \mathbf{d}_5\}|} = \frac{1}{5} = 0.2$$

Υπό Συνθήκη Πιθανότητες (Posterior / Conditional)

- Η πιθανότητα ενός γεγονότος σε ένα πλαίσιο όπου γνωρίζουμε ότι ένα ή περισσότερα γεγονότα έχουν ήδη συμβεί.
- Το σύμβολο $|$ διαβάζεται ως *δεδομένου ότι*.
- Το πλήθος των γραμμών όπου είναι αληθή και τα δύο γεγονότα, διαιρεμένο με το πλήθος των γραμμών όπου είναι αληθές μόνο το «δεδομένο» γεγονός.

Example

$$P(\text{Die1} = \text{'6'} \mid \text{Die2} = \text{'5'}) = \frac{|\{\mathbf{d}_3\}|}{|\{\mathbf{d}_2, \mathbf{d}_3\}|} = \frac{1}{2} = 0.5$$

Πίνακας: Ένα απλό dataset για το Meningitis με τρία συνηθισμένα συμπτώματα της νόσου ως περιγραφικά features: Headache, Fever και Vomiting.

ID	Headache	Fever	Vomiting	Meningitis
11	True	True	False	False
37	False	True	False	False
42	True	False	True	False
49	True	False	True	False
54	False	True	False	True
57	True	False	True	False
73	True	False	True	False
75	True	False	True	True
89	False	True	False	False
92	True	False	True	True

Η σειρά σου!

ID	Headach	Fever	Vomit	Meningitis
11	True	True	False	False
37	False	True	False	False
42	True	False	True	False
49	True	False	True	False
54	False	True	False	True
57	True	False	True	False
73	True	False	True	False
75	True	False	True	True
89	False	True	False	False
92	True	False	True	True

- $P(h) = ?$
- $P(m|h) = ?$
- $P(m, h) = ?$

Η σειρά σου!

$$P(h) = \frac{|\{\mathbf{d}_{11}, \mathbf{d}_{42}, \mathbf{d}_{49}, \mathbf{d}_{57}, \mathbf{d}_{73}, \mathbf{d}_{75}, \mathbf{d}_{92}\}|}{|\{\mathbf{d}_{11}, \mathbf{d}_{37}, \mathbf{d}_{42}, \mathbf{d}_{49}, \mathbf{d}_{54}, \mathbf{d}_{57}, \mathbf{d}_{73}, \mathbf{d}_{75}, \mathbf{d}_{89}, \mathbf{d}_{92}\}|} = \frac{7}{10} = 0.7$$

$$P(m|h) = \frac{|\{\mathbf{d}_{75}, \mathbf{d}_{92}\}|}{|\{\mathbf{d}_{11}, \mathbf{d}_{42}, \mathbf{d}_{49}, \mathbf{d}_{57}, \mathbf{d}_{73}, \mathbf{d}_{75}, \mathbf{d}_{92}\}|} = \frac{2}{7} = 0.2857$$

$$P(m, h) = \frac{|\{\mathbf{d}_{75}, \mathbf{d}_{92}\}|}{|\{\mathbf{d}_{11}, \mathbf{d}_{37}, \mathbf{d}_{42}, \mathbf{d}_{49}, \mathbf{d}_{54}, \mathbf{d}_{57}, \mathbf{d}_{73}, \mathbf{d}_{75}, \mathbf{d}_{89}, \mathbf{d}_{92}\}|} = \frac{2}{10} = 0.2$$

Κατανομές Πιθανότητας (Probability Distributions)

- Μια κατανομή πιθανότητας είναι μια δομή δεδομένων που περιγράφει, για όλες τις τιμές στο πεδίο τιμών ενός feature, την πιθανότητα το feature να πάρει κάθε τιμή.
- Μια κατανομή πιθανότητας ενός κατηγορικού feature είναι ένα διάνυσμα που περιέχει τις πιθανότητες που αντιστοιχούν στις τιμές του πεδίου τιμών του feature.
- Χρησιμοποιούμε έντονη γραφή $\mathbf{P}()$ για να ξεχωρίζουμε ότι μιλάμε για **κατανομή πιθανότητας** και όχι για μεμονωμένη τιμή από μια probability mass function $P()$.

Example

- Με βάση το παρακάτω dataset, η κατανομή πιθανότητας για το δυαδικό feature Meningitis, με τη σύμβαση ότι το πρώτο στοιχείο του διανύσματος είναι η πιθανότητα για *'True'*, είναι:

ID	Headach	Fever	Vomit	Meningitis
11	True	True	False	False
37	False	True	False	False
42	True	False	True	False
49	True	False	True	False
54	False	True	False	True
57	True	False	True	False
73	True	False	True	False
75	True	False	True	True
89	False	True	False	False
92	True	False	True	True

$$\mathbf{P}(M) = < 0.3, 0.7 >$$

Κοινές Κατανομές Πιθανότητας (Joint Probability Distributions)

- Μια κοινή κατανομή πιθανότητας είναι ένας πολυδιάστατος πίνακας, όπου κάθε κελί περιέχει την πιθανότητα για ένα συγκεκριμένο γεγονός στον δειγματικό χώρο που ορίζεται από τον συνδυασμό τιμών των features.

Example

- Η κοινή κατανομή πιθανότητας για τα τέσσερα δυαδικά features Headache, Fever, Vomiting, Meningitis στο πεδίο του Meningitis θα είναι:

$$\mathbf{P}(H, F, V, M) = \begin{bmatrix} P(h,f,v,m), & P(\neg h,f,v,m) \\ P(h,f,v,\neg m), & P(\neg h,f,v,\neg m) \\ P(h,f,\neg v,m), & P(\neg h,f,\neg v,m) \\ P(h,f,\neg v,\neg m), & P(\neg h,f,\neg v,\neg m) \\ P(h,\neg f,v,m), & P(\neg h,\neg f,v,m) \\ P(h,\neg f,v,\neg m), & P(\neg h,\neg f,v,\neg m) \\ P(h,\neg f,\neg v,m), & P(\neg h,\neg f,\neg v,m) \\ P(h,\neg f,\neg v,\neg m), & P(\neg h,\neg f,\neg v,\neg m) \end{bmatrix}$$

- Μια **πλήρης κοινή κατανομή πιθανότητας** (full joint probability distribution) είναι απλώς μια κοινή κατανομή πιθανότητας πάνω σε *όλα* τα features ενός πεδίου.

Άθροιση / Εξάλειψη Μεταβλητών (Summing out / Marginalisation)

- Δεδομένης μιας πλήρους κοινής πιθανότητας, μπορούμε να υπολογίσουμε την πιθανότητα οποιουδήποτε γεγονότος στο πεδίο αθροίζοντας τα κελιά της κοινής κατανομής στα οποία το γεγονός αυτό είναι αληθές.

Example

- Ας υποθέσουμε ότι θέλουμε να υπολογίσουμε την πιθανότητα $P(h)$ στο πεδίο που ορίζεται από την κοινή κατανομή $\mathbf{P}(H, F, V, M)$.
- Απλώς αθροίζουμε τις τιμές στα κελιά όπου ισχύει το h, δηλαδή τα κελιά της πρώτης στήλης.

$$\mathbf{P}(H, F, V, M) = \begin{bmatrix} P(h, f, v, m), & P(\neg h, f, v, m) \\ P(h, f, v, \neg m), & P(\neg h, f, v, \neg m) \\ P(h, f, \neg v, m), & P(\neg h, f, \neg v, m) \\ P(h, f, \neg v, \neg m), & P(\neg h, f, \neg v, \neg m) \\ P(h, \neg f, v, m), & P(\neg h, \neg f, v, m) \\ P(h, \neg f, v, \neg m), & P(\neg h, \neg f, v, \neg m) \\ P(h, \neg f, \neg v, m), & P(\neg h, \neg f, \neg v, m) \\ P(h, \neg f, \neg v, \neg m), & P(\neg h, \neg f, \neg v, \neg m) \end{bmatrix}$$

- Μπορούμε επίσης να χρησιμοποιήσουμε το summing out για να υπολογίσουμε **κοινές πιθανότητες** από μια κοινή κατανομή πιθανότητας.

Example

- Ας υποθέσουμε ότι θέλουμε να υπολογίσουμε την πιθανότητα των h και f, όταν δεν μας ενδιαφέρει τι τιμές παίρνουν τα V και M (εδώ τα V και M είναι παραδείγματα ενός κρυφού feature (hidden feature): feature του οποίου η τιμή δεν καθορίζεται ως evidence και δεν είναι target feature).

Example

- Για να υπολογίσουμε $P(h, V=?, M=?, f)$ από την $\mathbf{P}(H, F, V, M)$, αθροίζουμε τις τιμές σε όλα τα κελιά όπου ισχύουν τα h και f (δηλαδή αθροίζουμε τα τέσσερα πάνω κελιά της πρώτης στήλης).

$$\mathbf{P}(H, F, V, M) = \begin{bmatrix} P(h, f, v, m), & P(\neg h, f, v, m) \\ P(h, f, v, \neg m), & P(\neg h, f, v, \neg m) \\ P(h, f, \neg v, m), & P(\neg h, f, \neg v, m) \\ P(h, f, \neg v, \neg m), & P(\neg h, f, \neg v, \neg m) \\ P(h, \neg f, v, m), & P(\neg h, \neg f, v, m) \\ P(h, \neg f, v, \neg m), & P(\neg h, \neg f, v, \neg m) \\ P(h, \neg f, \neg v, m), & P(\neg h, \neg f, \neg v, m) \\ P(h, \neg f, \neg v, \neg m), & P(\neg h, \neg f, \neg v, \neg m) \end{bmatrix}$$

Υπό Συνθήκη Πιθανότητα (Conditional Probability)

$$P(X|Y) = \frac{P(X, Y)}{P(Y)} \tag{1}$$

- Χρησιμοποίησε αυτόν τον κανόνα για να ξαναϋπολογίσεις την πιθανότητα $P(m|h)$ (θυμήσου ότι $P(h) = 0.7$ και $P(m, h) = 0.2$).

Υπό Συνθήκη Πιθανότητα (Conditional Probability)

$$P(X|Y) = \frac{P(X,Y)}{P(Y)} \tag{1}$$

- Χρησιμοποίησε αυτόν τον κανόνα για να ξαναϋπολογίσεις την πιθανότητα $P(m|h)$ (θυμήσου ότι $P(h) = 0.7$ και $P(m,h) = 0.2$).

Example

$$P(m|h) = \frac{P(m,h)}{P(h)} = \frac{0.2}{0.7} = 0.2857$$

Κανόνας Γινομένου (Product Rule)

$$P(X, Y) = P(X|Y) \times P(Y)$$

- Σημείωση: $P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)$

Example

- Χρησιμοποίησε τον κανόνα γινομένου για να ξαναϋπολογίσεις το $P(m, h)$.

Κανόνας Γινομένου (Product Rule)

$$P(X, Y) = P(X|Y) \times P(Y)$$

- Σημείωση: $P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)$

Example

- Χρησιμοποίησε τον κανόνα γινομένου για να ξαναϋπολογίσεις το $P(m, h)$.

$$P(m, h) = P(m|h) \times P(h) = 0.2857 \times 0.7 = 0.2$$

Κανόνας Αλυσίδας (Chain Rule)

- Ο κανόνας γινομένου:

$$P(X, Y) = P(X|Y) \times P(Y)$$

- γενικεύεται στον κανόνα αλυσίδας:

$$\begin{aligned} P(A, B, C, \dots, Z) = & P(Z) \times P(Y|Z) \times P(X|Y, Z) \times \dots \\ & \times P(A|B, \dots, X, Y, Z) \end{aligned}$$

Θεώρημα Ολικής Πιθανότητας (Theorem of Total Probability)

$$P(X) = \sum_{i=1}^{k} P(X|Y_i)P(Y_i)$$

Example

- Χρησιμοποίησε το Θεώρημα Ολικής Πιθανότητας για να ξαναϋπολογίσεις το $P(h)$ αθροίζοντας (summing out) το M.

$$P(h) = (P(h|m) \times P(m)) + (P(h|\neg m) \times P(\neg m))$$

ID	Headach	Fever	Vomit	Meningitis
11	True	True	False	False
37	False	True	False	False
42	True	False	True	False
49	True	False	True	False
54	False	True	False	True
57	True	False	True	False
73	True	False	True	False
75	True	False	True	True
89	False	True	False	False
92	True	False	True	True

- $P(h|m) = ?$
- $P(m) = ?$
- $P(h|\neg m) = ?$
- $P(\neg m) = ?$

ID	Headach	Fever	Vomit	Meningitis
11	True	True	False	False
37	False	True	False	False
42	True	False	True	False
49	True	False	True	False
54	False	True	False	True
57	True	False	True	False
73	True	False	True	False
75	True	False	True	True
89	False	True	False	False
92	True	False	True	True

- $P(h|m) = 0.6666$
- $P(m) = 0.3$
- $P(h|\neg m) = 0.7143$
- $P(\neg m) = 0.7$

$$
\begin{aligned}
P(h) &= (P(h|m) \times P(m)) + (P(h|\neg m) \times P(\neg m)) \\
&= ?
\end{aligned}
$$

$$
\begin{aligned}
P(h) &= (P(h|m) \times P(m)) + (P(h|\neg m) \times P(\neg m)) \\
&= (0.6666 \times 0.3) + (0.7143 \times 0.7) = 0.7
\end{aligned}
$$

- Μπορούμε, αν θέλουμε, να αθροίσουμε (sum out) περισσότερα από ένα features.

Example

- Για παράδειγμα, μπορούμε να υπολογίσουμε το $P(h)$ αθροίζοντας όλα τα άλλα features στο dataset:

$$P(h) = \sum_{i \in level(M)} \sum_{j \in level(Fev)} \sum_{k \in level(V)} P(h|M_i, Fev_j, V_k) \times P(M_i, Fev_j, V_k)$$