Μάθηση βασισμένη στην πληροφορία

1 Βασική Ιδέα

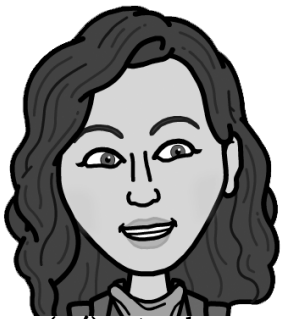
2 Θεμελιώδεις Έννοιες
- Δέντρα Αποφάσεων
- Μοντέλο Εντροπίας του Shannon
- Κέρδος Πληροφορίας

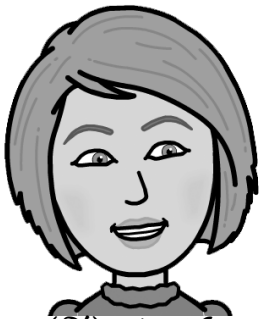
- Σε αυτό το κεφάλαιο θα παρουσιάσουμε έναν αλγόριθμο μηχανικής μάθησης που προσπαθεί να κατασκευάσει προγνωστικά μοντέλα **χρησιμοποιώντας μόνο τα πιο πληροφοριακά χαρακτηριστικά**.
- Σε αυτό το πλαίσιο, ένα πληροφοριακό χαρακτηριστικό είναι ένα **περιγραφικό χαρακτηριστικό** του οποίου οι τιμές χωρίζουν τα παραδείγματα του συνόλου δεδομένων σε **ομοιογενή σύνολα** ως προς την τιμή του χαρακτηριστικού-στόχου.

Βασική Ιδέα


(α΄) Brian


(β΄) John


(γ΄) Aphra

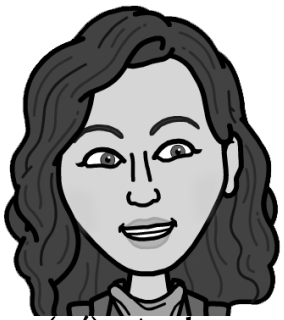

(δ΄) Aoife

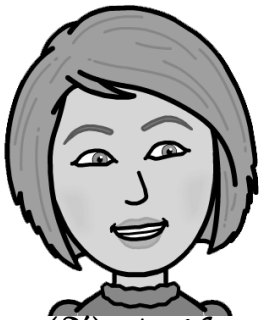
Σχήμα: Κάρτες που δείχνουν πρόσωπα και ονόματα χαρακτήρων για το παιχνίδι *Guess-Who* .

Άνδρας	**Μακριά Μαλλιά**	**Γυαλιά**	**Όνομα**
Ναι	Όχι	Ναι	Brian
Ναι	Όχι	Όχι	John
Όχι	Ναι	Όχι	Aphra
Όχι	Όχι	Όχι	Aoife


(α΄) Brian


(β΄) John

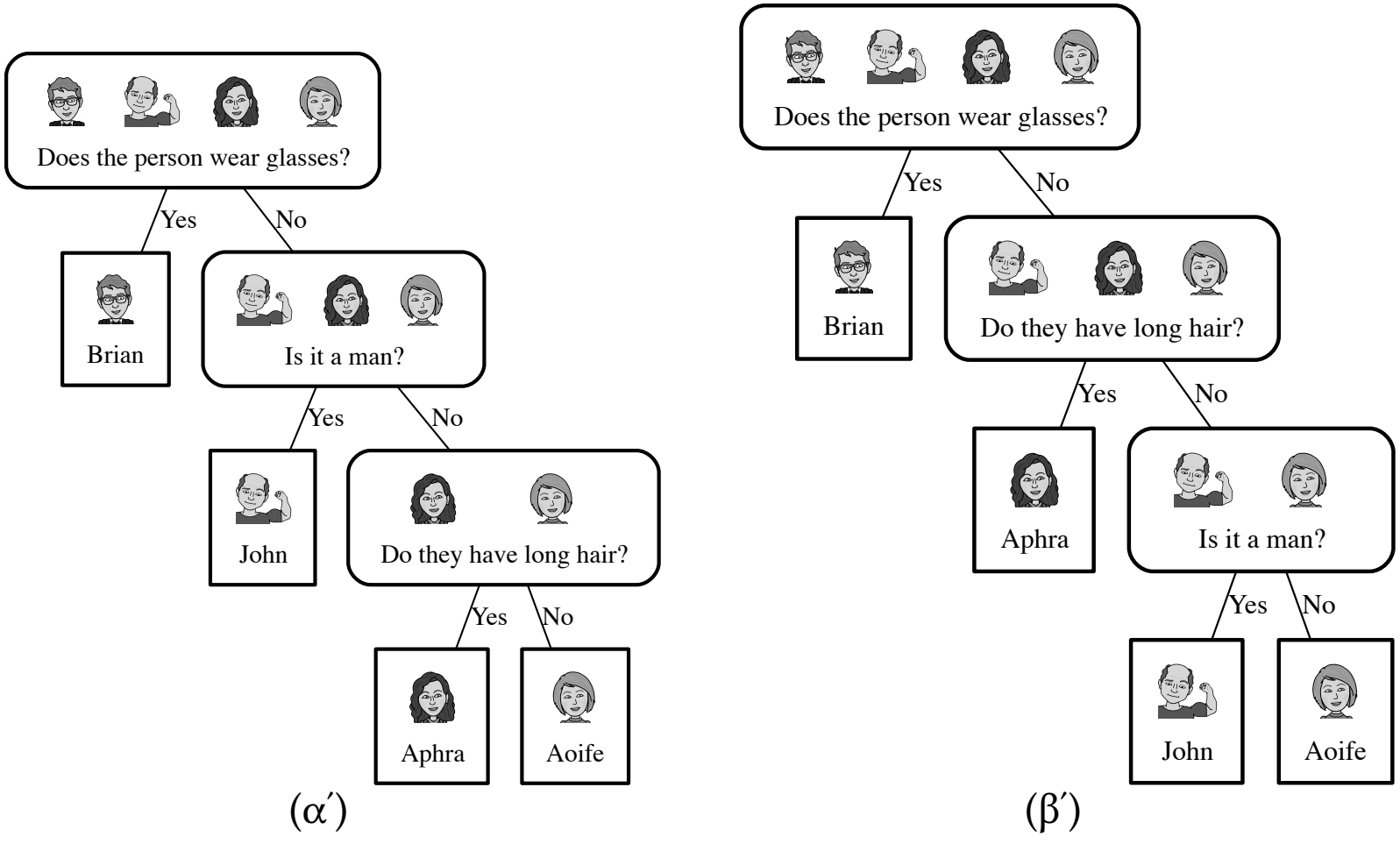

(γ΄) Aphra


(δ΄) Aoife

Σχήμα: Κάρτες που δείχνουν πρόσωπα και ονόματα χαρακτήρων για το παιχνίδι *Guess-Who* .

Ποια ερώτηση θα κάνατε πρώτη:

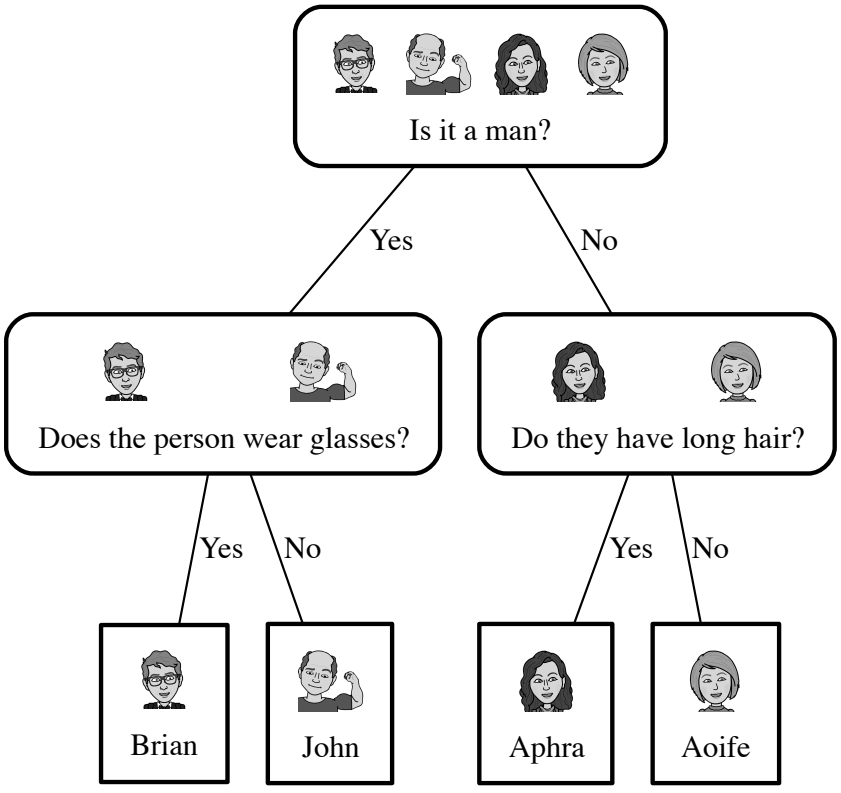
1. Είναι άνδρας;
2. Φοράει γυαλιά;



Σχήμα: Οι διαφορετικές ακολουθίες ερωτήσεων που μπορούν να ακολουθήσουν σε ένα παιχνίδι *Guess-Who* όταν ξεκινάμε με την ερώτηση Φοράει γυαλιά;

- Και στα δύο διαγράμματα:
  - ένα μονοπάτι έχει μήκος 1 ερώτηση,
  - ένα μονοπάτι έχει μήκος 2 ερωτήσεις,
  - και δύο μονοπάτια έχουν μήκος 3 ερωτήσεις.
- Επομένως, αν κάνετε πρώτα την Ερώτηση (2), ο μέσος αριθμός ερωτήσεων ανά παιχνίδι είναι:

$$\frac{1+2+3+3}{4} = 2.25$$



Σχήμα: Οι διαφορετικές ακολουθίες ερωτήσεων που μπορούν να ακολουθήσουν σε ένα παιχνίδι *Guess-Who* όταν ξεκινάμε με την ερώτηση Είναι άνδρας;

- Όλα τα μονοπάτια σε αυτό το διάγραμμα έχουν μήκος δύο ερωτήσεων.
- Άρα, κατά μέσο όρο, αν κάνετε πρώτα την Ερώτηση (1), ο μέσος αριθμός ερωτήσεων ανά παιχνίδι είναι:

$$\frac{2+2+2+2}{4} = 2$$

- Κατά μέσο όρο, η απάντηση στην Ερώτηση (1) φαίνεται να δίνει περισσότερη πληροφορία από την απάντηση στην Ερώτηση (2): απαιτούνται λιγότερες επόμενες ερωτήσεις.
- Αυτό δεν οφείλεται στο κυριολεκτικό περιεχόμενο των απαντήσεων: ΝΑΙ ή ΟΧΙ.
- Σχετίζεται με το πώς η απάντηση σε κάθε ερώτηση χωρίζει το πεδίο σε σύνολα διαφορετικού μεγέθους με βάση την τιμή του περιγραφικού χαρακτηριστικού και την πιθανότητα κάθε πιθανής απάντησης.

Βασική Ιδέα

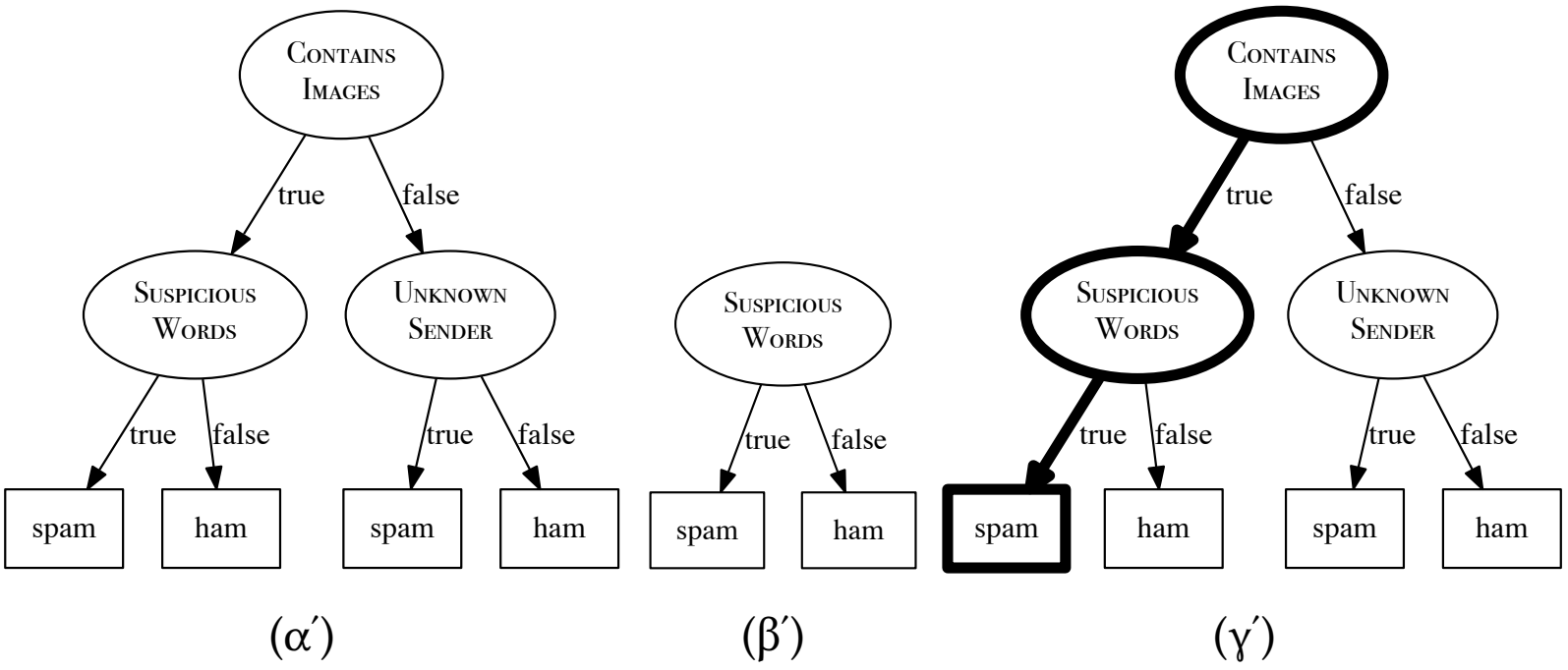
- Η βασική ιδέα εδώ είναι να εντοπίσουμε ποια χαρακτηριστικά είναι τα πιο πληροφοριακά για να θέτουμε ερωτήσεις γύρω από αυτά, εξετάζοντας τις επιδράσεις των διαφορετικών απαντήσεων ως προς:
  1. το πώς χωρίζεται το πεδίο μετά την απάντηση,
  2. και την πιθανότητα καθεμιάς από τις απαντήσεις.

Θεμελιώδεις Έννοιες

- Ένα δέντρο αποφάσεων αποτελείται από:
  1. έναν **ριζικό κόμβο** (ή αρχικό κόμβο),
  2. **εσωτερικούς κόμβους**,
  3. και **κόμβους φύλλων** (ή τερματικούς κόμβους).
- Κάθε μη φύλλο του δέντρου (ρίζα και εσωτερικοί κόμβοι) ορίζει έναν έλεγχο πάνω σε ένα από τα περιγραφικά χαρακτηριστικά του ερωτήματος.
- Κάθε κόμβος φύλλου ορίζει μια προβλεπόμενη κατηγορία για το ερώτημα.

Πίνακας: Ένα σύνολο δεδομένων για πρόβλεψη ανεπιθύμητης αλληλογραφίας email.

ID	Ύποπτες Λέξεις	Άγνωστος Αποστολέας	Περιέχει Εικόνες	Κλάση
376	true	false	true	spam
489	true	true	false	spam
541	true	true	false	spam
693	false	true	true	ham
782	false	false	false	ham
976	false	false	false	ham



Σχήμα: Τα (a) και (b) δείχνουν δύο δέντρα αποφάσεων που είναι συνεπή με τα παραδείγματα του συνόλου δεδομένων spam. Το (c) δείχνει τη διαδρομή που ακολουθείται στο δέντρο του (a) για να γίνει πρόβλεψη για το ερώτημα: Ύποπτες Λέξεις = *'true'*, Άγνωστος Αποστολέας = *'true'*, Περιέχει Εικόνες = *'true'*.

- Και τα δύο αυτά δέντρα θα επιστρέφουν ακριβώς τις ίδιες προβλέψεις για όλα τα παραδείγματα του συνόλου δεδομένων.
- Άρα, ποιο δέντρο πρέπει να χρησιμοποιήσουμε;

- Εφαρμόζουμε την ίδια προσέγγιση που χρησιμοποιήσαμε στο παιχνίδι *Guess-Who* : προτιμούμε δέντρα αποφάσεων που χρησιμοποιούν λιγότερους ελέγχους (ρηχότερα δέντρα).
- Αυτό είναι ένα παράδειγμα του ξυραφιού του Occam.

Πώς δημιουργούμε ρηχά δέντρα;

- Το δέντρο που ελέγχει το Ύποπτες Λέξεις στη ρίζα είναι πολύ ρηχό, επειδή το χαρακτηριστικό Ύποπτες Λέξεις χωρίζει τέλεια τα δεδομένα σε καθαρές ομάδες *'spam'* και *'ham'*.
- Τα περιγραφικά χαρακτηριστικά που χωρίζουν το σύνολο δεδομένων σε καθαρά σύνολα ως προς το χαρακτηριστικό-στόχο παρέχουν πληροφορία για το χαρακτηριστικό-στόχο.
- Επομένως, μπορούμε να δημιουργούμε ρηχά δέντρα ελέγχοντας τα πληροφοριακά χαρακτηριστικά νωρίς στο δέντρο.
- Το μόνο που χρειαζόμαστε για αυτό είναι ένα υπολογιστικό μέτρο καθαρότητας ενός συνόλου: η εντροπία

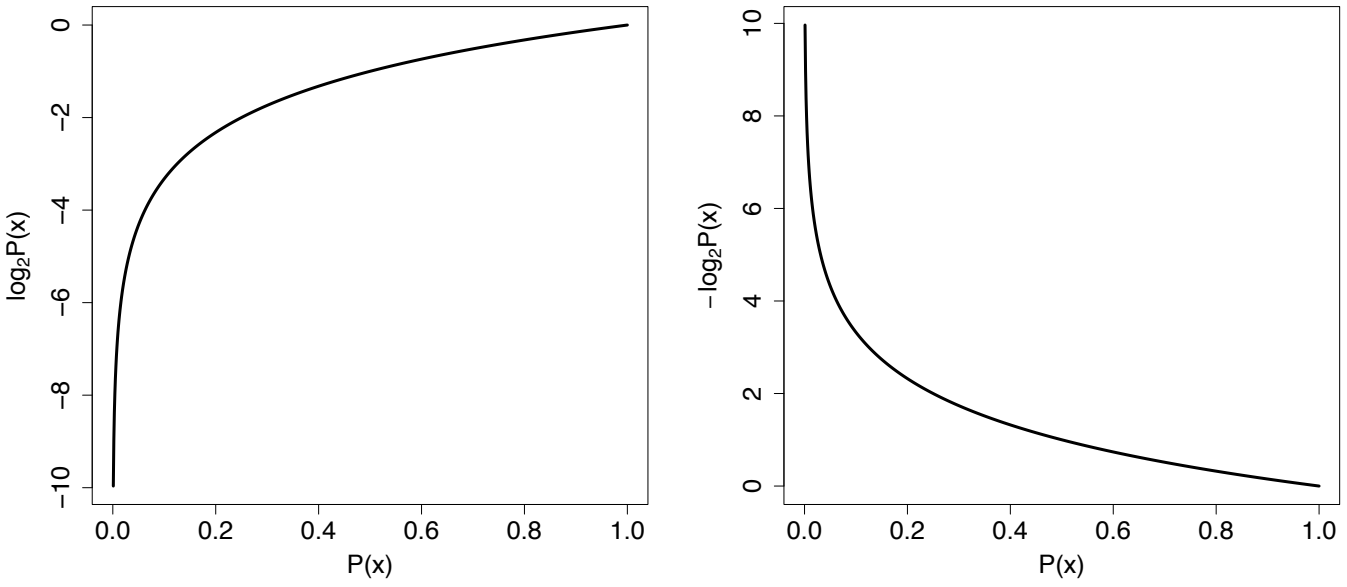
- Το μοντέλο εντροπίας του Claude Shannon ορίζει ένα υπολογιστικό μέτρο της ακαθαρσίας των στοιχείων ενός συνόλου.
- Ένας εύκολος τρόπος να κατανοήσουμε την εντροπία ενός συνόλου είναι να σκεφτούμε την αβεβαιότητα που σχετίζεται με την πρόβλεψη του αποτελέσματος όταν κάνουμε μια τυχαία επιλογή από το σύνολο.

- Η εντροπία σχετίζεται με την πιθανότητα ενός αποτελέσματος.
  - Υψηλή πιθανότητα $\rightarrow$ Χαμηλή εντροπία
  - Χαμηλή πιθανότητα $\rightarrow$ Υψηλή εντροπία
- Αν πάρουμε τον λογάριθμο μιας πιθανότητας και τον πολλαπλασιάσουμε με -1, παίρνουμε αυτή την αντιστοίχιση!

Τι είναι ο λογάριθμος;

Θυμηθείτε ότι ο λογάριθμος του a με βάση το b είναι ο αριθμός στον οποίο πρέπει να υψώσουμε το b ώστε να πάρουμε το a.

- $log_2(0.5) = -1$ επειδή $2^{-1} = 0.5$
- $log_2(1) = 0$ επειδή $2^0 = 1$
- $log_2(8) = 3$ επειδή $2^3 = 8$
- $log_5(25) = 2$ επειδή $5^2 = 25$
- $log_5(32) = 2.153$ επειδή $5^{2.153} = 32$



(α′) (β′)

Σχήμα: (a) Γράφημα που δείχνει πώς αλλάζει η τιμή του δυαδικού λογαρίθμου (λογάριθμος με βάση το 2) μιας πιθανότητας σε όλο το εύρος των τιμών πιθανότητας. (b) Η επίδραση του πολλαπλασιασμού αυτών των τιμών με -1.

- Το μοντέλο εντροπίας του Shannon είναι ένα σταθμισμένο άθροισμα των λογαρίθμων των πιθανοτήτων κάθε πιθανού αποτελέσματος όταν κάνουμε μια τυχαία επιλογή από ένα σύνολο.

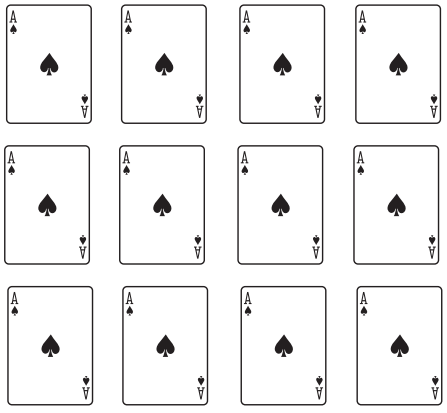
$$H(t) = -\sum_{i=1}^{l} (P(t=i) \times log_s(P(t=i))) \tag{1}$$

- Ποια είναι η εντροπία ενός συνόλου από 52 διαφορετικά τραπουλόχαρτα;

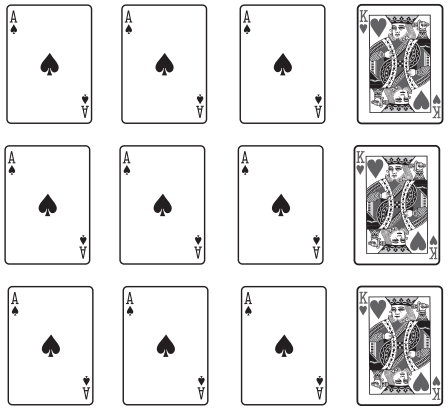
$$
\begin{aligned}
H(card) &= -\sum_{i=1}^{52} P(card = i) \times log_2(P(card = i)) \\
&= -\sum_{i=1}^{52} 0.019 \times log_2(0.019) = -\sum_{i=1}^{52} -0.1096 \\
&= 5.700 \; bits
\end{aligned}
$$

- Ποια είναι η εντροπία ενός συνόλου από 52 τραπουλόχαρτα αν τα διακρίνουμε μόνο με βάση το χρώμα τους $\{\heartsuit, \clubsuit, \diamondsuit, \spadesuit\}$;

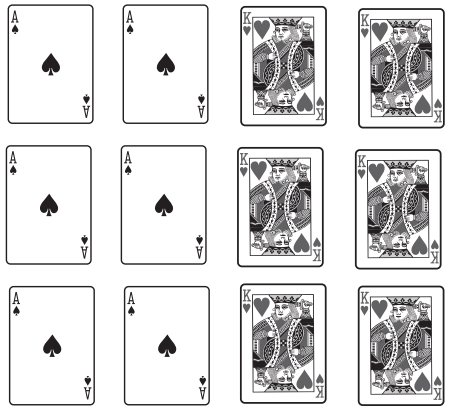
$$
\begin{aligned}
H(suit) &= -\sum_{l \in \{\heartsuit,\clubsuit,\diamondsuit,\spadesuit\}} P(suit = l) \times log_2(P(suit = l)) \\
&= -\big((P(\heartsuit) \times log_2(P(\heartsuit))) + (P(\clubsuit) \times log_2(P(\clubsuit))) \\
&\qquad + (P(\diamondsuit) \times log_2(P(\diamondsuit))) + (P(\spadesuit) \times log_2(P(\spadesuit))) \big) \\
&= -\big(\big({}^{13}/_{52} \times log_2({}^{13}/_{52})\big) + \big({}^{13}/_{52} \times log_2({}^{13}/_{52})\big) \\
&\qquad + \big({}^{13}/_{52} \times log_2({}^{13}/_{52})\big) + \big({}^{13}/_{52} \times log_2({}^{13}/_{52})\big) \big) \\
&= -\big((0.25 \times -2) + (0.25 \times -2) \\
&\qquad + (0.25 \times -2) + (0.25 \times -2) \big) \\
&= 2 \; bits
\end{aligned}
$$



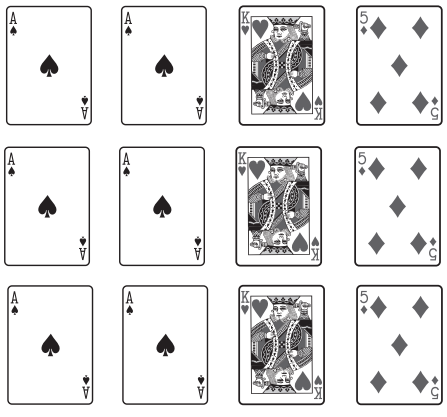
(α΄) $H(card) = 0.00$



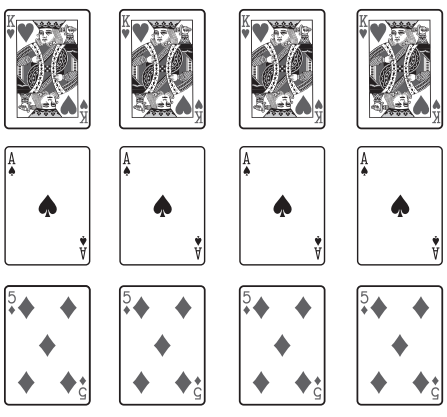
(β΄) $H(card) = 0.81$



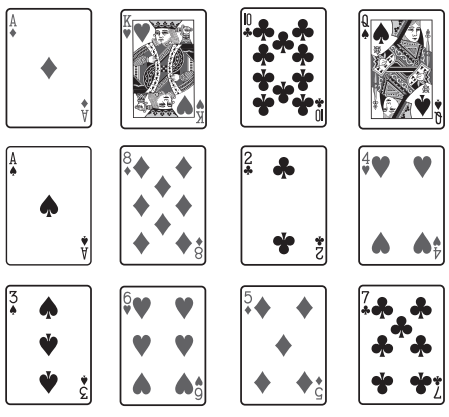
(γ΄) $H(card) = 1.00$



(δ΄) $H(card) = 1.50$



(ε΄) $H(card) = 1.58$

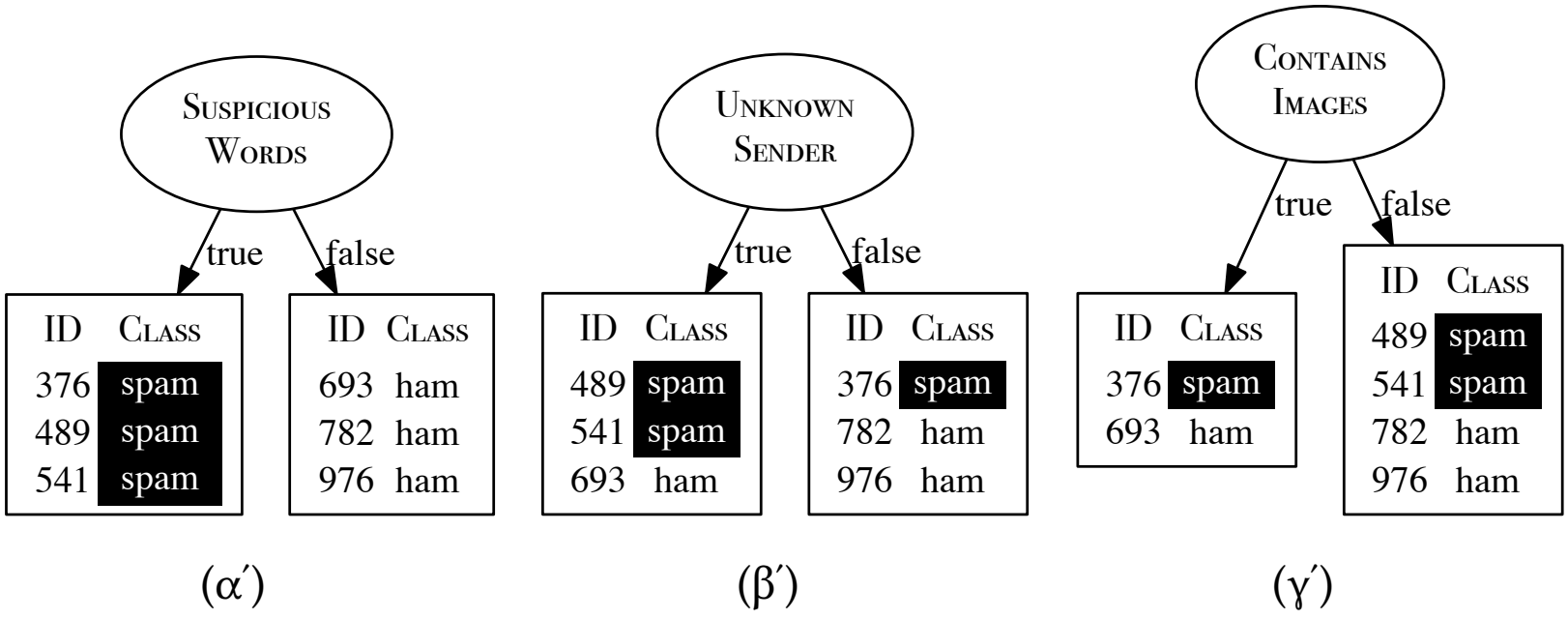


(ς΄) $H(card) = 3.58$

Σχήμα: Η εντροπία διαφορετικών συνόλων τραπουλόχαρτων, μετρημένη σε bits.

Πίνακας: Η σχέση μεταξύ της εντροπίας ενός μηνύματος και του συνόλου από το οποίο επιλέχθηκε.

Εντροπία Μηνύματος	**Ιδιότητες του Συνόλου Μηνυμάτων**
Υψηλή	Ένα μεγάλο σύνολο εξίσου πιθανών μηνυμάτων.
Μεσαία	Ένα μεγάλο σύνολο μηνυμάτων, μερικά πιο πιθανά από άλλα.
Μεσαία	Ένα μικρό σύνολο εξίσου πιθανών μηνυμάτων.
Χαμηλή	Ένα μικρό σύνολο μηνυμάτων με ένα πολύ πιθανό μήνυμα.



Σχήμα: Πώς χωρίζονται τα παραδείγματα του συνόλου δεδομένων spam όταν τα διαμερίζουμε με βάση καθένα από τα περιγραφικά χαρακτηριστικά του συνόλου δεδομένων spam στον Πίνακα 1 [15]

- Η διαίσθησή μας είναι ότι το ιδανικό διακριτικό χαρακτηριστικό θα διαμερίζει τα δεδομένα σε **καθαρά** υποσύνολα όπου όλα τα παραδείγματα σε κάθε υποσύνολο έχουν την ίδια κατηγορία.
  - Ύποπτες Λέξεις: τέλειος διαχωρισμός.
  - Άγνωστος Αποστολέας: μίγμα, αλλά με κάποια πληροφορία (όταν είναι *'true'*, τα περισσότερα παραδείγματα είναι *'spam'*).
  - Περιέχει Εικόνες: καμία πληροφορία.
- Ένας τρόπος να υλοποιήσουμε αυτή την ιδέα είναι να χρησιμοποιήσουμε ένα μέτρο που λέγεται **κέρδος πληροφορίας**.

Κέρδος Πληροφορίας

- Το κέρδος πληροφορίας ενός περιγραφικού χαρακτηριστικού μπορεί να ερμηνευτεί ως μέτρο της μείωσης της συνολικής εντροπίας ενός έργου πρόβλεψης όταν ελέγχουμε αυτό το χαρακτηριστικό.

Ο υπολογισμός του κέρδους πληροφορίας περιλαμβάνει τις ακόλουθες 3 εξισώσεις:

$$H(t, \mathcal{D}) = -\sum_{l \in levels(t)} (P(t = l) \times log_2(P(t = l))) \tag{2}$$

$$rem(d, \mathcal{D}) = \sum_{l \in levels(d)} \underbrace{\frac{|\mathcal{D}_{d=l}|}{|\mathcal{D}|}}_{\text{στάθμιση}} \times \underbrace{H(t, \mathcal{D}_{d=l})}_{\substack{\text{εντροπία του} \\ \text{διαμερίσματος } \mathcal{D}_{d=l}}} \tag{3}$$

$$IG(d, \mathcal{D}) = H(t, \mathcal{D}) - rem(d, \mathcal{D}) \tag{4}$$

- Ως παράδειγμα, θα υπολογίσουμε το κέρδος πληροφορίας για καθένα από τα περιγραφικά χαρακτηριστικά στο σύνολο δεδομένων spam email.

- Υπολογίστε την εντροπία για το χαρακτηριστικό-στόχο στο σύνολο δεδομένων.

$$H(t, \mathcal{D}) = - \sum_{l \in levels(t)} (P(t = l) \times log_2(P(t = l)))$$

ID	Ύποπτες Λέξεις	Άγνωστος Αποστολέας	Περιέχει Εικόνες	Κλάση
376	true	false	true	spam
489	true	true	false	spam
541	true	true	false	spam
693	false	true	true	ham
782	false	false	false	ham
976	false	false	false	ham

$$
\begin{aligned}
H(t, \mathcal{D}) &= - \sum_{l \in \{\text{'spam'}, \text{'ham'}\}} (P(t = l) \times log_2(P(t = l))) \\
&= - ((P(t = \text{'spam'}) \times log_2(P(t = \text{'spam'})) \\
&\quad + (P(t = \text{'ham'}) \times log_2(P(t = \text{'ham'}))) \\
&= -\left(\left({}^3/_6 \times log_2({}^3/_6)\right) + \left({}^3/_6 \times log_2({}^3/_6)\right)\right) \\
&= 1 \ bit
\end{aligned}
$$

- Υπολογίστε το υπόλοιπο για το χαρακτηριστικό Ύποπτες Λέξεις στο σύνολο δεδομένων.

$$rem\,(d, \mathcal{D}) = \sum_{l \in levels(d)} \underbrace{\frac{|\mathcal{D}_{d=l}|}{|\mathcal{D}|}}_{\text{στάθμιση}} \times \underbrace{H\,(t, \mathcal{D}_{d=l})}_{\substack{\text{εντροπία του} \\ \text{διαμερίσματος } \mathcal{D}_{d=l}}}$$

$$
\begin{aligned}
&rem\,(\textsc{words}, \mathcal{D}) \\
&\quad = \left(\frac{|\mathcal{D}_{\textsc{words}=T}|}{|\mathcal{D}|} \times H(t, \mathcal{D}_{\textsc{words}=T})\right) + \left(\frac{|\mathcal{D}_{\textsc{words}=F}|}{|\mathcal{D}|} \times H(t, \mathcal{D}_{\textsc{words}=F})\right) \\
&\quad = \left({}^{3}/_{6} \times \left(- \sum_{l \in \{\text{'spam'}, \text{'ham'}\}} P(t=l) \times log_2(P(t=l))\right)\right) \\
&\quad + \left({}^{3}/_{6} \times \left(- \sum_{l \in \{\text{'spam'}, \text{'ham'}\}} P(t=l) \times log_2(P(t=l))\right)\right) \\
&\quad = \left({}^{3}/_{6} \times \left(- \left(\left({}^{3}/_{3} \times log_2({}^{3}/_{3})\right) + \left({}^{0}/_{3} \times log_2({}^{0}/_{3})\right)\right)\right)\right) \\
&\quad + \left({}^{3}/_{6} \times \left(- \left(\left({}^{0}/_{3} \times log_2({}^{0}/_{3})\right) + \left({}^{3}/_{3} \times log_2({}^{3}/_{3})\right)\right)\right)\right) = 0\ bits
\end{aligned}
$$

- Υπολογίστε το υπόλοιπο για το χαρακτηριστικό Άγνωστος Αποστολέας στο σύνολο δεδομένων.

$$
\begin{aligned}
&rem\,(\textsc{sender}, \mathcal{D}) \\
&= \left(\frac{|\mathcal{D}_{\textsc{sender}=T}|}{|\mathcal{D}|} \times H(t, \mathcal{D}_{\textsc{sender}=T})\right) + \left(\frac{|\mathcal{D}_{\textsc{sender}=F}|}{|\mathcal{D}|} \times H(t, \mathcal{D}_{\textsc{sender}=F})\right) \\
&= \left({}^{3}/_{6} \times \left(- \sum_{l \in \{\text{'spam'}, \text{'ham'}\}} P(t=l) \times log_2(P(t=l))\right)\right) \\
&+ \left({}^{3}/_{6} \times \left(- \sum_{l \in \{\text{'spam'}, \text{'ham'}\}} P(t=l) \times log_2(P(t=l))\right)\right) \\
&= \left({}^{3}/_{6} \times \left(- \left(\left({}^{2}/_{3} \times log_2({}^{2}/_{3})\right) + \left({}^{1}/_{3} \times log_2({}^{1}/_{3})\right)\right)\right)\right) \\
&+ \left({}^{3}/_{6} \times \left(- \left(\left({}^{1}/_{3} \times log_2({}^{1}/_{3})\right) + \left({}^{2}/_{3} \times log_2({}^{2}/_{3})\right)\right)\right)\right) = 0.9183\ bits
\end{aligned}
$$

- Υπολογίστε το υπόλοιπο για το χαρακτηριστικό Περιέχει Εικόνες στο σύνολο δεδομένων.

$$
\begin{aligned}
&rem\left(\textsc{images}, \mathcal{D}\right) \\
&\quad = \left(\frac{|\mathcal{D}_{\textsc{images}=T}|}{|\mathcal{D}|} \times H\left(t, \mathcal{D}_{\textsc{images}=T}\right)\right) + \left(\frac{|\mathcal{D}_{\textsc{images}=F}|}{|\mathcal{D}|} \times H\left(t, \mathcal{D}_{\textsc{images}=F}\right)\right) \\
&\quad = \left({}^{2}/_{6} \times \left(- \sum_{l \in \{\text{`spam'}, \text{`ham'}\}} P(t=l) \times log_2(P(t=l))\right)\right) \\
&\quad + \left({}^{4}/_{6} \times \left(- \sum_{l \in \{\text{`spam'}, \text{`ham'}\}} P(t=l) \times log_2(P(t=l))\right)\right) \\
&\quad = \left({}^{2}/_{6} \times \left(- \left(\left({}^{1}/_{2} \times log_2({}^{1}/_{2})\right) + \left({}^{1}/_{2} \times log_2({}^{1}/_{2})\right)\right)\right)\right) \\
&\quad + \left({}^{4}/_{6} \times \left(- \left(\left({}^{2}/_{4} \times log_2({}^{2}/_{4})\right) + \left({}^{2}/_{4} \times log_2({}^{2}/_{4})\right)\right)\right)\right) = 1 \; bit
\end{aligned}
$$

- Υπολογίστε το κέρδος πληροφορίας για τα τρία περιγραφικά χαρακτηριστικά του συνόλου δεδομένων.

$$IG(d, \mathcal{D}) = H(t, \mathcal{D}) - rem(d, \mathcal{D})$$

$$\begin{aligned} IG(\textsc{Suspicious Words}, \mathcal{D}) &= H(\textsc{Class}, \mathcal{D}) - rem(\textsc{Suspicious Words}, \mathcal{D}) \\ &= 1 - 0 = 1 \ bit \end{aligned}$$

$$\begin{aligned} IG(\textsc{Unknown Sender}, \mathcal{D}) &= H(\textsc{Class}, \mathcal{D}) - rem(\textsc{Unknown Sender}, \mathcal{D}) \\ &= 1 - 0.9183 = 0.0817 \ bits \end{aligned}$$

$$\begin{aligned} IG(\textsc{Contains Images}, \mathcal{D}) &= H(\textsc{Class}, \mathcal{D}) - rem(\textsc{Contains Images}, \mathcal{D}) \\ &= 1 - 1 = 0 \ bits \end{aligned}$$

- Τα αποτελέσματα αυτών των υπολογισμών συμφωνούν με τη διαίσθησή μας.