Κατηγοριοποίηση και Πρόβλεψη

Δρ. Σωτήριος Δ. Νικολόπουλος

Big Data & Analytics

Πανεπιστήμιο Πελοποννήσου

Τμήμα Λογιστικής & Χρηματοοικονομικών

s.nikolopoulos@go.uop.gr

Συνοπτική Στατιστική και Οπτικοποίηση

1 Κατηγοριοποίηση και Πρόβλεψη

Κατηγοριοποίηση και Πρόβλεψη

1 Κατηγοριοποίηση και Πρόβλεψη

Κατηγοριοποίηση και Πρόβλεψη

Καιρός	Θερμοκρασία	Υγρασία	Αέρας	Class
Ηλιόλουστος	Υψηλή	Υψηλή	Ασθενής	Μέσα
Ηλιόλουστος	Υψηλή	Υψηλή	Δυνατός	Μέσα
Συννεφιά	Υψηλή	Υψηλή	Ασθενής	Έξω
Βροχή	Κανονική	Υψηλή	Ασθενής	Έξω
Βροχή	Χαμηλή	Κανονική	Δυνατός	Μέσα
Συννεφιά	Χαμηλή	Κανονική	Ασθενής	Έξω
Βροχή	Κανονική	Κανονική	Ασθενής	Έξω
Συννεφιά	Υψηλή	Κανονική	Ασθενής	Έξω

Κατηγοριοποίηση και Πρόβλεψη

Για να χωρίσουμε τα δεδομένα χρησιμοποιώντας την εντροπία και το κέρδος πληροφορίας, πρέπει πρώτα να υπολογίσουμε την εντροπία του αρχικού συνόλου δεδομένων και στη συνέχεια να υπολογίσουμε το κέρδος πληροφορίας για κάθε γνώρισμα (Καιρός, Θερμοκρασία, Υγρασία και Αέρα).

Θα επιλέξουμε το γνώρισμα με το υψηλότερο κέρδος πληροφορίας για να χωρίσουμε περαιτέρω τα δεδομένα.

Ας ξεκινήσουμε με τον υπολογισμό της εντροπίας του αρχικού συνόλου δεδομένων.

Η εντροπία είναι μια μέτρηση της ακαθαρσίας ή της τυχαιότητας ή της ανομοιογένειας στο σύνολο δεδομένων.

Κατηγοριοποίηση και Πρόβλεψη

Για ένα πρόβλημα δυαδικής ταξινόμησης όπως αυτό, η συνάρτηση εντροπίας $H(S)$ ορίζεται ως εξής:

$$H(S) = -p_1 \log_2(p_1) - p_2 \log_2(p_2)$$

Όπου

p_1 είναι το ποσοστό των παραδειγμάτων/περιπτώσεων της κλάσης 1, και

p_2 είναι το ποσοστό των παραδειγμάτων/περιπτώσεων της κλάσης 2.

Από τα δεδομένα του παραδείγματος:

Υπάρχουν 3 παραδείγματα της κατηγορίας ”Μέσα” και 5 παραδείγματα της κατηγορίας ”Έξω”.

Άρα:

$$p_{\text{Μέσα}} = \frac{3}{8} \quad \text{και} \quad p_{\text{Έξω}} = \frac{5}{8}$$

$$H(S) = -\left(\frac{3}{8}\log_2\frac{3}{8} + \frac{5}{8}\log_2\frac{5}{8}\right)$$

$$H(S) = -\left(\frac{3}{8}\times(-1.415) + \frac{5}{8}\times(-0.678)\right)$$

$$H(S) = 0.954$$

Κατηγοριοποίηση και Πρόβλεψη

Τώρα, θα υπολογίσουμε το κέρδος πληροφορίας για κάθε γνώρισμα (Καιρός, Θερμοκρασία, Υγρασία και Αέρα), διαχωρίζοντας το σύνολο δεδομένων με βάση κάθε γνώρισμα και υπολογίζοντας την εντροπία κάθε διαχωρισμού.

Ας ξεκινήσουμε με το γνώρισμα του Καιρού:

Καιρός:
- **Ηλιόλουστος:**
- 2 Περιπτώσεις (Μέσα, Μέσα)
- **Συννεφιά:**
- 3 Περιπτώσεις (Έξω, Έξω, Έξω)
- **Βροχή:**
- 3 Περιπτώσεις (Έξω, Μέσα, Έξω)

$$H(\text{Ηλιόλουστος}) = -\left[\frac{2}{2}\log_2\left(\frac{2}{2}\right) + \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right] = 0$$

$$H(\text{Συννεφιά}) = -\left[\frac{3}{3}\log_2\left(\frac{3}{3}\right) + \frac{0}{3}\log_2\left(\frac{0}{3}\right)\right] = 0$$

$$H(\text{Βροχή}) = -\left[\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right] = 0.918$$

Τώρα, υπολόγισε το κέρδος πληροφορίας:

$$IG(\text{Καιρός}) = H(S) - \left(\frac{2}{8}\times 0 + \frac{3}{8}\times 0 + \frac{3}{8}\times 0.918\right)$$

$$IG(\text{Καιρός}) = 0.954 - 0.344$$

$$IG(\text{Καιρός}) = 0.61$$

Κατηγοριοποίηση και Πρόβλεψη

Επαναλαμβάνουμε την ίδια διαδικασία για τα γνωρίσματα Θερμοκρασίας, Υγρασίας και Αέρα και επιλέγουμε αυτό με το υψηλότερο κέρδος πληροφορίας για να διαχωρίσουμε περαιτέρω τα δεδομένα.

Ας συνεχίσουμε με τον υπολογισμό του κέρδους πληροφορίας για τα γνωρίσματα Θερμοκρασίας, Υγρασίας και Αέρα.

Κατηγοριοποίηση και Πρόβλεψη

Θερμοκρασία:

- **Υψηλή**:

- 4 Περιπτώσεις (Μέσα, Μέσα, Έξω, Έξω)

- **Κανονική**:

- 2 Περιπτώσεις (Έξω, Έξω)

- **Χαμηλή**:

- 2 Περιπτώσεις (Μέσα, Έξω)

$$H(\text{Υψηλή}) = -\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) = 1$$

$$H(\text{Κανονική}) = -\left(\frac{0}{2}\log_2\frac{0}{2} + \frac{2}{2}\log_2\frac{2}{2}\right) = 0$$

$$H(\text{Χαμηλή}) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$$

Τώρα, υπολόγισε το κέρδος πληροφορίας:

$$IG(\text{Θερμοκρασία}) = H(S) - \left(\frac{4}{8} \times 1 + \frac{2}{8} \times 0 + \frac{2}{8} \times 1\right)$$

$$IG(\text{Θερμοκρασία}) = 0.954 - 0.750$$

$$IG(\text{Θερμοκρασία}) = 0.204$$

Υγρασία:

- ****Υψηλή****:

- 4 Περιπτώσεις (Μέσα, Μέσα, Έξω, Έξω)

- ****Κανονική****:

- 4 Περιπτώσεις (Μέσα, Έξω, Έξω, Έξω)

$$H(\text{Υψηλή}) = -\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) = 1$$

$$H(\text{Κανονική}) = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}\right) = 0.811$$

Τώρα, υπολόγισε το κέρδος πληροφορίας:

$$IG(\text{Υγρασία}) = H(S) - \left(\frac{4}{8}\times 1 + \frac{4}{8}\times 0.811\right)$$

$$IG(\text{Υγρασία}) = 0.954 - 0.906 = 0.048$$

Υγρασία:

- ****Υψηλή****:

- 4 Περιπτώσεις (Μέσα, Μέσα, Έξω, Έξω)

- ****Κανονική****:

- 4 Περιπτώσεις (Μέσα, Έξω, Έξω, Έξω)

$$H(\text{Υψηλή}) = -\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{4}{4}\log_2\frac{4}{4}\right) = 1$$

$$H(\text{Κανονική}) = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}\right) = 0.811$$

Τώρα, υπολόγισε το κέρδος πληροφορίας:

$$IG(\text{Υγρασία}) = H(S) - \left(\frac{4}{8}\times 1 + \frac{4}{8}\times 0.811\right)$$

$$IG(\text{Υγρασία}) = 0.954 - 0.906 = 0.048$$

Αέρας:

- **Ασθενής:**

- 6 Περιπτώσεις (Μέσα, Έξω, Έξω, Έξω, Έξω, Έξω)

- **Δυνατός:**

- 2 Περιπτώσεις (Μέσα, Μέσα)

$$H(\text{Ασθενής}) = -\left(\frac{1}{6}\log_2\frac{1}{6} + \frac{5}{6}\log_2\frac{5}{6}\right) = 0.650$$

$$H(\text{Δυνατός}) = -\left(\frac{2}{2}\log_2\frac{2}{2} + \frac{0}{2}\log_2\frac{0}{2}\right) = 0$$

Τώρα, υπολόγισε το κέρδος πληροφορίας:

$$IG(\text{Αέρας}) = H(S) - \left(\frac{6}{8}\times 0.650 + \frac{2}{8}\times 0\right)$$

$$IG(\text{Αέρας}) = 0.954 - 0.488 = 0.466$$

Κατηγοριοποίηση και Πρόβλεψη

Σύνοψη των Κερδών Πληροφορίας:

- Καιρός: 0.61
- Θερμοκρασία: 0.204
- Υγρασία: 0.048
- Αέρας: = 0.466

Το γνώρισμα με το υψηλότερο κέρδος πληροφορίας είναι **Καιρός.**

Συνεπώς, θα πρέπει να διαχωρίσουμε το σύνολο δεδομένων με βάση το γνώρισμα του Καιρού.

Κατηγοριοποίηση και Πρόβλεψη

Όταν ο καιρός είναι Ηλιοφάνεια, η κλάση είναι πάντα Μέσα
Όταν ο καιρός είναι Συννεφιά, η κλάση είναι πάντα Έξω
Όταν ο καιρός είναι Βροχή, η κλάση μπορεί να είναι μέσα ή έξω

Άρα όταν ο καιρός είναι Ηλιοφάνεια ή Συννεφιά δεν μπορεί να γίνει διαχωρισμός

Όταν ο καιρός είναι Βροχή μπορεί να γίνει περαιτέρω διάσπαση.

Για τον σκοπό αυτό δημιουργούμε ένα Υποσύνολο όπου ο καιρός είναι Βροχή

Για να υπολογίσουμε την εντροπία του Υποσυνόλου 1 (Καιρός = Βροχή) και στη συνέχεια να υπολογίσουμε τα κέρδη πληροφορίας για τα άλλα γνωρίσματα εντός του Υποσυνόλου 1, θα ακολουθήσουμε τα ίδια βήματα όπως πριν.

Κατηγοριοποίηση και Πρόβλεψη

Υποσύνολο 1: Καιρός = Βροχή

Καιρός	**Θερμοκρασία**	**Υγρασία**	**Αέρας**	**Κατηγορία**
Βροχή	Κανονική	Υψηλή	Αδύναμος	Έξω
Βροχή	Χαμηλή	Κανονική	Δυνατός	Μέσα
Βροχή	Κανονική	Κανονική	Αδύναμος	Έξω

Ας υπολογίσουμε πρώτα την εντροπία του Υποσυνόλου 1:

1. Μετρήστε τις εμφανίσεις κάθε κατηγορίας στο Υποσύνολο 1:

Κατηγορία	**Μέτρηση**
Μέσα	1
Έξω	2

2. Υπολογίστε την εντροπία χρησιμοποιώντας αυτές τις μετρήσεις:

$$H(\text{Υποσύνολο 1}) = -\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right) \approx \mathbf{0.918}$$

Τώρα, ας υπολογίσουμε τα κέρδη πληροφορίας για τα άλλα γνωρίσματα (Θερμοκρασία, Υγρασία και Αέρας) εντός του Υποσυνόλου 1:

Θερμοκρασία:

- **Κανονική:**

- 2 Περιπτώσεις (Έξω, Έξω)

- **Χαμηλή:**

- 1 Περίπτωση (Μέσα)

$$H(\text{Κανονική}) = -\left(\frac{2}{2}\log_2\left(\frac{2}{2}\right) + \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right) = 0$$

$$H(\text{Χαμηλή}) = -\left(\frac{0}{1}\log_2\frac{0}{1} + \frac{1}{1}\log_2\frac{1}{1}\right) = 0$$

Τώρα, υπολογίστε το κέρδος πληροφορίας:

$$IG(\text{Θερμοκρασία}) = H(\text{Υποσύνολο 1}) - \left(\frac{1}{3} \times 0 + \frac{2}{3} \times 0\right)$$

$$IG(\text{Θερμοκρασία}) = 0.918 - 0$$

$$IG(\text{Θερμοκρασία}) = \mathbf{0.918}$$

Υγρασία:

- ****Υψηλή****:

- 1 Περίπτωση (Έξω)

- ****Κανονική****:

- 2 Περιπτώσεις (Μέσα, Έξω)

$$H(\text{Υψηλή}) = -\left(\frac{0}{1}\log_2\frac{0}{1} + \frac{1}{1}\log_2\frac{1}{1}\right) = 0$$

$$H(\text{Κανονική}) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$$

Τώρα, υπολογίστε το κέρδος πληροφορίας:

$$IG(\text{Υγρασία}) = H(\text{Υποσύνολο 1}) - \left(\frac{1}{3}\times 0 + \frac{2}{3}\times 1\right)$$

$$IG(\text{Υγρασία}) = 0.918 - \frac{2}{3} \approx 0.252$$

Αέρας:

- ****Αδύναμος**:**

- 2 Περιπτώσεις (Έξω, Έξω)

- **Δυνατός**:

- 1 Περίπτωση (Μέσα)

$$H(\text{Αδύναμος}) = -\left(\frac{2}{2}\log_2\frac{2}{2}\right) = 0$$

$$H(\text{Δυνατός}) = -\left(\frac{0}{1}\log_2\frac{0}{1} + \frac{1}{1}\log_2\frac{1}{1}\right) = 0$$

Τώρα, υπολογίστε το κέρδος πληροφορίας:

$$IG(\text{Αέρας}) = H(\text{Υποσύνολο 1}) - \left(\frac{2}{3}\times 0 + \frac{1}{3}\times 0\right)$$

$$IG(\text{Αέρας}) = 0.918 - 0$$

$$IG(\text{Αέρας}) = \mathbf{0.918}$$

Κατηγοριοποίηση και Πρόβλεψη

Επομένως, τα κέρδη πληροφορίας για τα γνωρίσματα Θερμοκρασίας, Υγρασίας και Αέρα εντός του Υποσυνόλου 1 (Καιρός = Βροχή) είναι περίπου:

- Θερμοκρασία: **0.918**
- Υγρασία: 0.252
- Αέρας: **0.918**