Train In Data

ANOVA

Evaluating continuous features

Leave length	Plant species
28	Species A
25.7	Species A
28.2	Species A
32.3	Species A
27.5	Species A
21.8	Species B
24	Species B
26.7	Species B
25.6	Species B
23.8	Species B
22.2	Species C
21.4	Species C
22.1	Species C
28.2	Species C
19.5	Species C

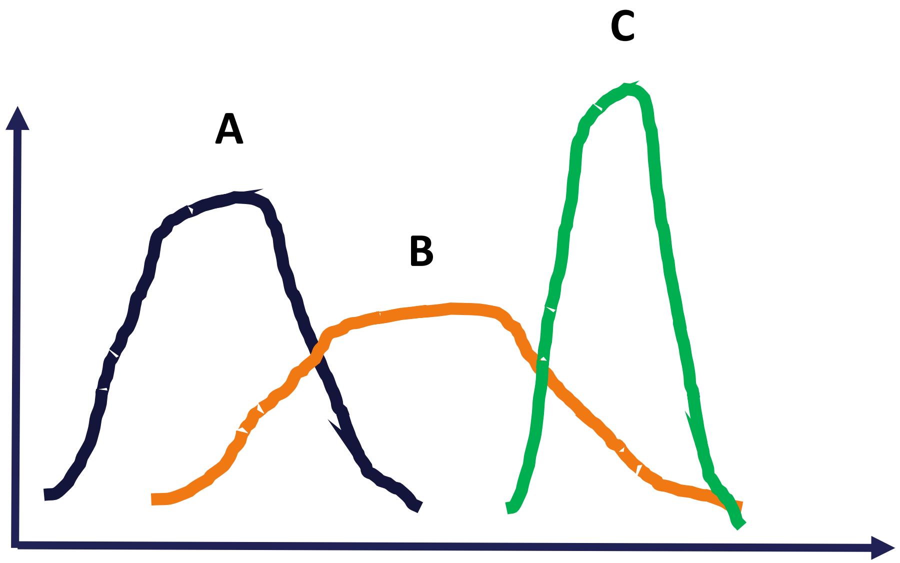
Continuous predictor, categorical target.

How can we know if *leave length* is predictive of *species*?

Evaluating continuous features

Leave length	Plant species
28	Species A
25.7	Species A
28.2	Species A
32.3	Species A
27.5	Species A
21.8	Species B
24	Species B
26.7	Species B
25.6	Species B
23.8	Species B
22.2	Species C
21.4	Species C
22.1	Species C
28.2	Species C
19.5	Species C

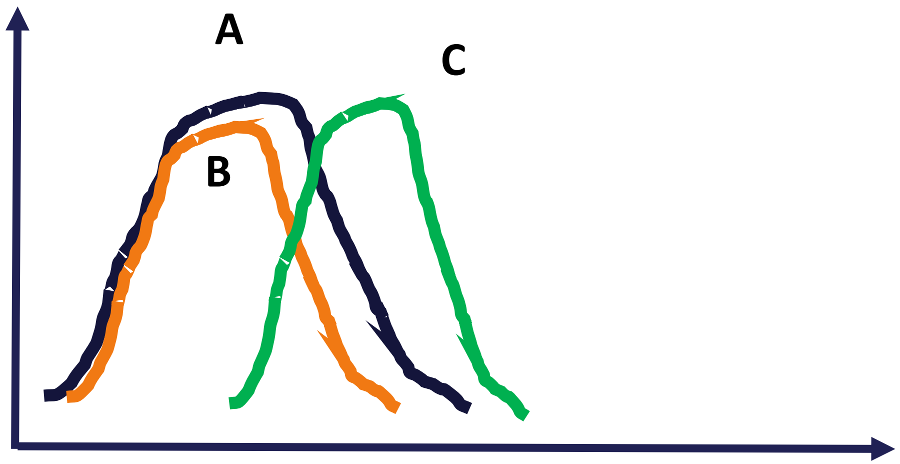
If the variable is predictive of species, we expect different distributions across species.



Evaluating continuous features

Leave length	Plant species
28	Species A
25.7	Species A
28.2	Species A
32.3	Species A
27.5	Species A
21.8	Species B
24	Species B
26.7	Species B
25.6	Species B
23.8	Species B
22.2	Species C
21.4	Species C
22.1	Species C
28.2	Species C
19.5	Species C

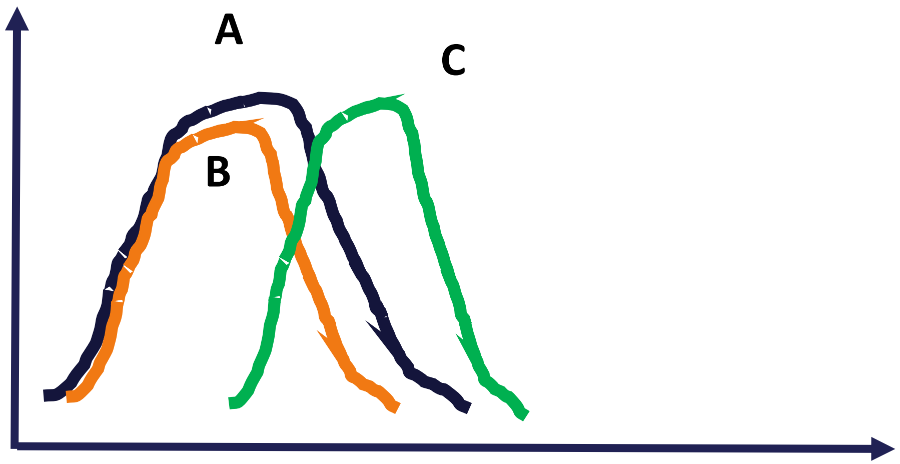
If the variable is not predictive of species, the distributions would overlap, to some degree.



Evaluating continuous features

Leave length	Plant species
28	Species A
25.7	Species A
28.2	Species A
32.3	Species A
27.5	Species A
21.8	Species B
24	Species B
26.7	Species B
25.6	Species B
23.8	Species B
22.2	Species C
21.4	Species C
22.1	Species C
28.2	Species C
19.5	Species C

How can we assess, statistically, if 2 or more distributions are the same?



ANOVA

ANOVA stands for Analysis Of Variance.

ANOVA tests if the mean of different groups come from the same population.

Considering the variance.

ANOVA assumptions

Anova tests the hypothesis that 2 or more samples have the same mean.

- Samples are independent
- Samples are normally distributed
- Homogeneity of variance

ANOVA

Leave length	Plant species
28	Species A
25.7	Species A
28.2	Species A
32.3	Species A
27.5	Species A
21.8	Species B
24	Species B
26.7	Species B
25.6	Species B
23.8	Species B
22.2	Species C
21.4	Species C
22.1	Species C
28.2	Species C
19.5	Species C

ANOVA decomposes the total variability of the data into "explained" and "unexplained".

ANOVA

Leave length	Plant species
28	Species A
25.7	Species A
28.2	Species A
32.3	Species A
27.5	Species A
21.8	Species B
24	Species B
26.7	Species B
25.6	Species B
23.8	Species B
22.2	Species C
21.4	Species C
22.1	Species C
28.2	Species C
19.5	Species C

ANOVA decomposes the total variability of the data into “explained” and “unexplained”.

- Total variability ➔ total sum of squares
- Explained variability ➔ model sum of squares
- Unexplained variability ➔ residual sum of squares

Total sum of squares

Leave length	Plant species
28	Species A
25.7	Species A
28.2	Species A
32.3	Species A
27.5	Species A
21.8	Species B
24	Species B
26.7	Species B
25.6	Species B
23.8	Species B
22.2	Species C
21.4	Species C
22.1	Species C
28.2	Species C
19.5	Species C

Total variability ➔ total sum of squares ➔ variance of the variable

$$SS_T = \sum(x_{ij} - \overline{x}_{grand})^2$$

Total sum of squares

Leave length	Plant species
28	Species A
25.7	Species A
28.2	Species A
32.3	Species A
27.5	Species A
21.8	Species B
24	Species B
26.7	Species B
25.6	Species B
23.8	Species B
22.2	Species C
21.4	Species C
22.1	Species C
28.2	Species C
19.5	Species C

Variable mean = 25.13

$$SS_T = \sum(x_{ij} - \overline{x}_{grand}))^2$$

$$SS_T = (28-25.13)^2 + (25.7-25.13)^2 + (28.2-25.13)^2 + (32.3-25.13)^2 + (27.5-25.13)^2 +$$
$$(21.8-25.13)^2 + (24-25.13)^2 + (26.7-25.13)^2 + (25.6-25.13)^2 + (23.8-25.13)^2 +$$
$$(22.2-25.13)^2 + (21.4-25.13)^2 + (22.1-25.13)^2 + (28.2-25.13)^2 + (19.5-25.13)^2$$

SST = 164.63

Model sum of squares

Leave length	Plant species
28	Species A
25.7	Species A
28.2	Species A
32.3	Species A
27.5	Species A
21.8	Species B
24	Species B
26.7	Species B
25.6	Species B
23.8	Species B
22.2	Species C
21.4	Species C
22.1	Species C
28.2	Species C
19.5	Species C

Some of the variability could be explained by the fact that different samples come from different groups.

$$SS_M = \sum n_k(\overline{x_j} - \overline{x}_{grand})^2$$

Model sum of squares

Leave length	Plant species
28	Species A
25.7	Species A
28.2	Species A
32.3	Species A
27.5	Species A
21.8	Species B
24	Species B
26.7	Species B
25.6	Species B
23.8	Species B
22.2	Species C
21.4	Species C
22.1	Species C
28.2	Species C
19.5	Species C

Some of the variability could be explained by the fact that different samples come from different groups.

$$SS_M = \sum n_k(\overline{x_j} - \overline{x}_{grand})^2$$

$$SS_M = 5(28.34 - 25.13)^2 + 5(24.38 - 25.13)^2 + 5(22.68 - 25.13)^2$$

SSM = 84.34

Residual sum of squares

Leave length	Plant species
28	Species A
25.7	Species A
28.2	Species A
32.3	Species A
27.5	Species A
21.8	Species B
24	Species B
26.7	Species B
25.6	Species B
23.8	Species B
22.2	Species C
21.4	Species C
22.1	Species C
28.2	Species C
19.5	Species C

The "unexplained" variability is the variability within groups.

$$SS_R = \sum (x_{ij} + \overline{x}_j)^2$$

Residual sum of squares

Leave length	Plant species
28	Species A
25.7	Species A
28.2	Species A
32.3	Species A
27.5	Species A
21.8	Species B
24	Species B
26.7	Species B
25.6	Species B
23.8	Species B
22.2	Species C
21.4	Species C
22.1	Species C
28.2	Species C
19.5	Species C

The "unexplained" variability is the variability within groups.

$$SS_R = \sum(x_{ij} - \bar{x}_j)^2$$

$$SS_R = (28 - 28.34)^2 + (25.7 - 28.34)^2 + (28.2 - 28.34)^2 + (32.3 - 28.34)^2 + (27.5 - 28.34)^2 +$$
$$(21.8 - 24.38)^2 + (24 - 24.38)^2 + (26.7 - 24.38)^2 + (25.6 - 24.38)^2 + (23.8 - 24.38)^2 +$$
$$(22.2 - 22.68)^2 + (21.4 - 22.68)^2 + (22.1 - 22.68)^2 + (28.2 - 22.68)^2 + (19.5 - 22.68)^2$$

SSR = 80.29

Degrees of freedom

Leave length	Plant species
28	Species A
25.7	Species A
28.2	Species A
32.3	Species A
27.5	Species A
21.8	Species B
24	Species B
26.7	Species B
25.6	Species B
23.8	Species B
22.2	Species C
21.4	Species C
22.1	Species C
28.2	Species C
19.5	Species C

We need to go from sum of squares to mean squares ➔ divide by degrees of freedom

- Dof SST = # samples -1 = 15 – 1 = 14
- Dof SSM = # groups -1 = 3 – 1 = 2
- Dof SSR = # samples - # groups = 15 – 3 = 12

Mean sum of squares

Leave length	Plant species
28	Species A
25.7	Species A
28.2	Species A
32.3	Species A
27.5	Species A
21.8	Species B
24	Species B
26.7	Species B
25.6	Species B
23.8	Species B
22.2	Species C
21.4	Species C
22.1	Species C
28.2	Species C
19.5	Species C

Mean variation explained and unexplained.

- MSM = ${}^{SSM}/_{Dof\ SSM}$ = ${}^{84.34}/_{2}$ = 42.17
- MSR = ${}^{SSR}/_{Dof\ SSR}$ = ${}^{80.29}/_{12}$ = 6.69

F-ratio

Leave length	Plant species
28	Species A
25.7	Species A
28.2	Species A
32.3	Species A
27.5	Species A
21.8	Species B
24	Species B
26.7	Species B
25.6	Species B
23.8	Species B
22.2	Species C
21.4	Species C
22.1	Species C
28.2	Species C
19.5	Species C

If the model can't explain any variability ➔ MSM is small.

Larger MSM ➔ more variability is explained by the model.

- $F = {}^{MSM}/_{MSR} = {}^{42.17}/_{6.69} = 6.30$

F-ratio

Leave length	Plant species
28	Species A
25.7	Species A
28.2	Species A
32.3	Species A
27.5	Species A
21.8	Species B
24	Species B
26.7	Species B
25.6	Species B
23.8	Species B
22.2	Species C
21.4	Species C
22.1	Species C
28.2	Species C
19.5	Species C

If the model does not explain any variability, F ➔ close to 1

Larger F ➔ more variability is explained by the model.

- F = $^{MSM}/_{MSR}$ = $^{42.17}/_{6.69}$ = 6.30

F-ratio

- F follows a well-known distribution that depends on the degrees of freedom of numerator and denominator.
- Knowing F ➔ p-value ➔ probability of 2 samples coming from the same distribution
- Smaller p-values ➔ distributions are different across groups.

ANOVA assumptions

- Samples are independent
- Samples are normally distributed
- Homogeneity of variance

When assumptions are not met ➔ variance stabilizing transformations (log, power, Box-Cox, etc)

ANOVA considerations

Effect size → with big data even small differences seem significant (small p-value).

Good for ranking features, but for statistical reliability we need to consider the size effect.

Anova: Scikit-learn

- `f_classif`: returns F and p-values.
  - Rank features
  - Larger F or smaller p-values ➔ important features
- `SelectKBest:` select best k features
- `SelectPercentile:` select features in top percentile

THANK YOU

www.trainindata.com