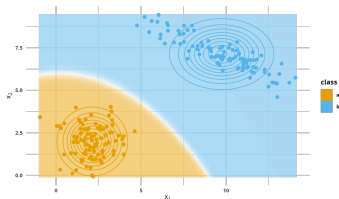


Introduction to Machine Learning

Classification

Naive Bayes



Learning goals

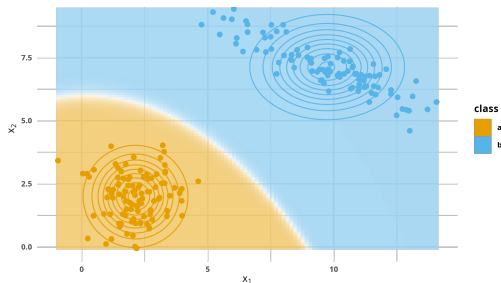
- Construction principle of NB
- Conditional independence assumption
- Numerical and categorical features
- Similarity to QDA, quadratic decision boundaries
- Laplace smoothing

NUMERICAL FEATURES

Use univariate Gaussians for $p(x_j|y = k)$, and estimate $(\mu_{kj}, \sigma_{kj}^2)$.

Because of $p(\mathbf{x}|y = k) = \prod_{j=1}^p p(x_j|y = k)$, joint conditional density is

Gaussian with diagonal, non-isotropic covariances, and different across classes, so **QDA with diagonal covariances**.



Note: In the above plot the data violates the NB assumption.

NB: CATEGORICAL FEATURES

We use a categorical distribution for $p(x_j|y = k)$ and estimate the probabilities p_{kjm} that, in class k , our j -th feature has value m , $x_j = m$, simply by counting frequencies.

$$p(x_j|y = k) = \prod_m p_{kjm}^{[x_j=m]}$$

Because of the simple conditional independence structure, it is also very easy to deal with mixed numerical / categorical feature spaces.

ID	Class	Sex	Survived the Titanic
1	2nd	male	no
2	1st	male	yes
3	3rd	female	yes
4	1st	female	yes
5	2nd	female	yes
6	3rd	female	no

$$p(x_{\text{sex}} | y = \text{yes}) = p_{\text{yes,sex,female}}^{[x_{\text{sex}}=\text{female}]} \cdot p_{\text{yes,sex,male}}^{[x_{\text{sex}}=\text{male}]}$$

$$= \frac{3}{4} \cdot \frac{1}{4}$$



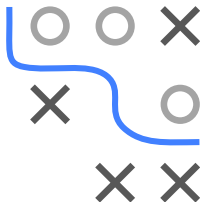
LAPLACE SMOOTHING

If a given class and feature value never occur together in the training data, then the frequency-based probability estimate will be zero, e.g.:

$$p_{\text{no, class, 1st}}^{[x_{\text{class}}=1\text{st}]} = 0 \text{ (everyone from 1st class survived in the previous table)}$$

This is problematic because it will wipe out all information in the other probabilities when they are multiplied!

$$\pi_{\text{no}}(\text{class} = 1\text{st}, \text{sex} = \text{male}) = \frac{\hat{p}(x_{\text{class}}|y = \text{no}) \cdot \hat{p}(x_{\text{sex}}|y = \text{no}) \cdot \hat{\pi}_{\text{no}}}{\sum_{j=1}^g \hat{p}(\text{class} = 1\text{st}, \text{sex} = \text{male}|y = j) \hat{\pi}_j} = 0$$



NAIVE BAYES: APPLICATION AS SPAM FILTER

- In the late 90s, NB became popular for e-mail spam detection
- Word counts were used as features to detect spam mails
- Independence assumption implies: occurrence of two words in mail is not correlated, this is often wrong;
"viagra" more likely to occur in context with "buy"...
- In practice: often still good performance



Benchmarking QDA, NB and LDA on spam:

