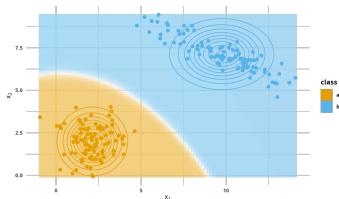
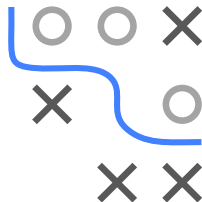


Introduction to Machine Learning

Classification

Naive Bayes



Learning goals

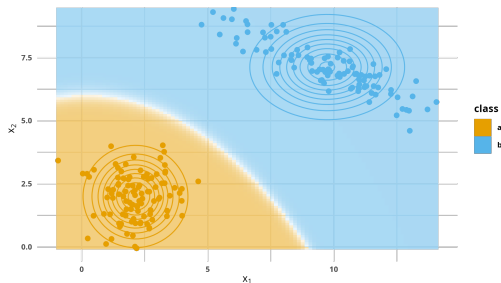
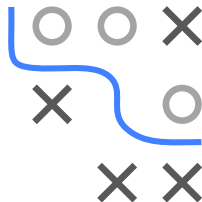
- Construction principle of NB
- Conditional independence assumption
- Numerical and categorical features
- Similarity to QDA, quadratic decision boundaries
- Laplace smoothing

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**.

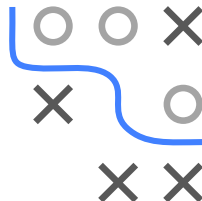


Note: In the above plot the data violates the NB assumption.

NB: CATEGORICAL FEATURES

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

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



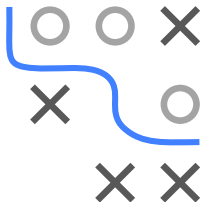
LAPLACE SMOOTHING

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.:

$p_{no, class, 1st}^{[x_{class}=1st]} = 0$ (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!

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



LAPLACE SMOOTHING

A simple numerical correction is to set these zero probabilities to a small value to regularize against this case.

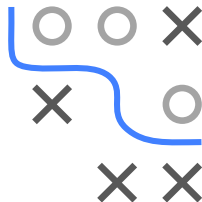
- Add constant $\alpha > 0$ (e.g., $\alpha = 1$).
- For a categorical feature x_j with M_j possible values:

$$p_{kjm}^{[x_j=m]} = \frac{n_{kjm} + \alpha}{n_k + \alpha M_j} \quad \left(\text{instead of } p_{kjm}^{[x_j=m]} = \frac{n_{kjm}}{n_k} \right)$$

where:

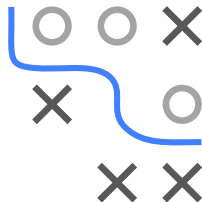
- n_{kjm} : count of $x_j = m$ in class k ,
- n_k : total counts in class k ,
- M_j : number of possible distinct values of x_j .

This ensures that our posterior probabilities are non-zero due to such effects, preserving the influence of all features in the model.



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:

