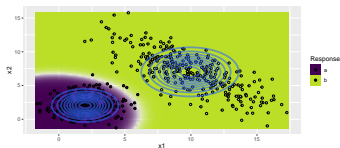


# Introduction to Machine Learning

## Classification

### Naive Bayes



### Learning goals

- Understand the idea of Naive Bayes
- Understand in which sense Naive Bayes is a special QDA model

# NAIVE BAYES CLASSIFIER

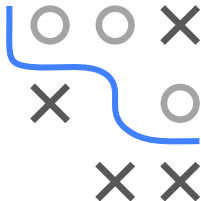
NB is a generative multiclass technique. Remember: We use Bayes' theorem and only need  $p(\mathbf{x}|y = k)$  to compute the posterior as:

$$\pi_k(\mathbf{x}) = \mathbb{P}(y = k \mid \mathbf{x}) = \frac{\mathbb{P}(\mathbf{x}|y = k)\mathbb{P}(y = k)}{\mathbb{P}(\mathbf{x})} = \frac{p(\mathbf{x}|y = k)\pi_k}{\sum_{j=1}^g p(\mathbf{x}|y = j)\pi_j}$$

NB is based on a simple **conditional independence assumption**: the features are conditionally independent given class  $y$ .

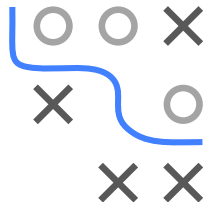
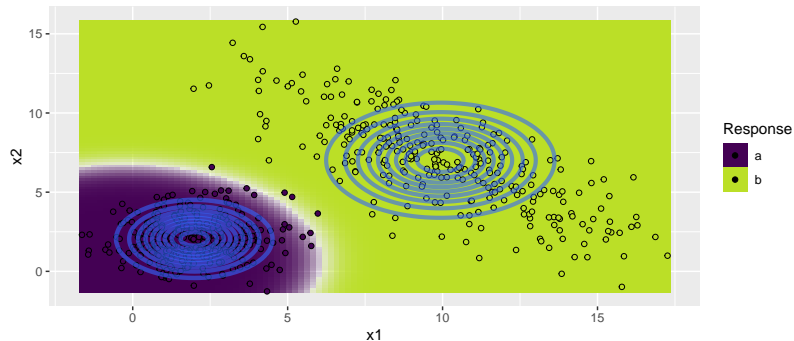
$$p(\mathbf{x}|y = k) = p((x_1, x_2, \dots, x_p)|y = k) = \prod_{j=1}^p p(x_j|y = k).$$

So we only need to specify and estimate the distribution  $p(x_j|y = k)$ , which is considerably simpler as this is univariate.



# NB: NUMERICAL FEATURES

We use a univariate Gaussian for  $p(x_j|y = k)$ , and estimate  $(\mu_j, \sigma_j^2)$  in the standard manner. Because of  $p(\mathbf{x}|y = k) = \prod_{j=1}^p p(x_j|y = k)$ , the joint conditional density is Gaussian with diagonal but non-isotropic covariance structure, and potentially different across classes. Hence, NB is a (specific) QDA model, with quadratic decision boundary.



## 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 the 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.



# 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.

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

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



# NAIVE BAYES: APPLICATION AS SPAM FILTER

- In the late 90s, Naive Bayes became popular for e-mail spam filter programs
- Word counts were used as features to detect spam mails (e.g., "Viagra" often occurs in spam mail)
- Independence assumption implies: occurrence of two words in mail is not correlated
- Seems naive ("Viagra" more likely to occur in context with "Buy now" than "flower"), but leads to less required parameters and therefore better generalization, and often works well in practice.

