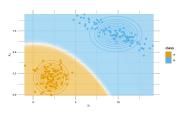
Introduction to Machine Learning

Classification Naive Bayes





Learning goals

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

NAIVE BAYES CLASSIFIER

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

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

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

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

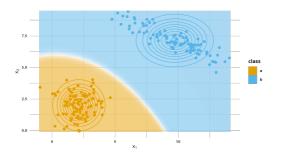
So we only need to specify and estimate the distributions $p(x_j|y = k)$, which is considerably simpler as these are univariate.

 \mathbf{X}

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.



× 0 0 × × ×

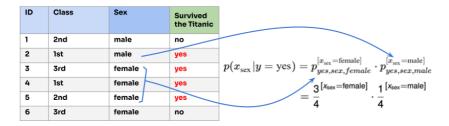
Note: In the above plot the data violates the NB assumption.

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.



× × 0 × × ×

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

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

× > 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_i with M_i possible values:

$$p_{kjm}^{[x_j=m]} = rac{n_{kjm} + lpha}{n_k + lpha M_j} \quad \left(ext{instead of } p_{kjm}^{[x_j=m]} = rac{n_{kjm}}{n_k}
ight)$$

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*.

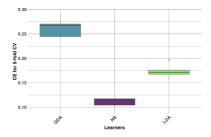
This ensures that our posterior probabilities are non-zero due to such effects, preserving the influence of all features in the model.



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:



× 0 0 × 0 × ×