Naive Bayes Classifier

The naive Bayes approach makes the assumption that all attributes are independant

This assumption means that the likelihood can be decomposed into a product of dimension-wise probabilities:

$$P(x|c_i) = P(x_1, x_2, \ldots, x_d | c_i) = \prod_{j=1}^{d} P(x_j | c_i)$$

How is this different?

- A full Bayes trys to model the joint Probability of all the features
- This means it takes into account the dependancies & correlations between each feature
- This is what makes Bayes complex & costly
- in theory, with enough Data it is more accurate as it captures true relationships

How A Naive Bayes is different?

- features are conditionally indep given class
- That is the value of one does not affect the value of another, given Ci
- This drastically reduces complexity
- Instead of est a complex joint distrib
- Naive Bayes only est marginal probs of each ~~class~~ feature given class
- Computationally efficient, req less train
- Generally still performs well partic in domains where independ ASSumption is valid

Naive Bayer: Numeric Attribs

Default assumption that each is normally distributed

let μ (mean), $\&$ σ^2 (var) be for attribute X_j for class C_i

Note here X_j is a single attribute but for all attributes

The likelihood for C_i given Dimension X_j

$$P(X_j|C_i) \propto f(X_j|\mu_{ij},\sigma_{ij}^2) = \underbrace{\frac{1}{\sqrt{2\pi}\,\sigma_{ij}} \exp\left\{-\frac{(x_j-\mu_{ij})^2}{2\sigma_{ij}^2}\right\}}_{\text{Prob Dense function}}$$

$\propto$ = Proportional given constant

tells us that the conditional probability is proportional to the gaussian PDF upto a constant factor

why? Bayes classifier we often care more about the relative probabilities of a data point belong to different classes rather than the exact probs

Addition consequence of the independance assumption corresponds to setting all covarances to 0 — (non-diags) Σ_i

$$\Sigma_i = \begin{pmatrix} \sigma_{i,1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{i2}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & \sigma_{id}^2 \end{pmatrix}$$

Indep so no cross interacts

Determinant of Σ_i: $(|\Sigma_i|)$

- the determinant of a matrix is a scalar value that provides info about mat
- for diag mats, it is the product of the diag elements
- This is one of the factors that makes the naive calc alot quicker overall

$$|\Sigma_i| = \det(\Sigma_i) = \sigma_{i1}^2 \cdot \sigma_{i2}^2 \cdot \ldots \cdot \sigma_{id}^2 = \prod_{j=1}^{d} \sigma_{ij}^2$$

Product of all vals

Inverse Matrix of covariance matrix

Σ_i^{-1} = Diag Matrix where all values are $\frac{1}{\sigma}$

we want the inverse because we can plug it into the Multivariate Gaussian PDF (likelihood)

$$(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i) \quad \sum_{j=1}^{d} \frac{(x_j-\mu_{ij})^2}{\sigma_{ij}^2}$$

Use d dimensions to account for all data Attribs

RHS of PDF

~~Rows of a data~~

Note a univariate PDF only looks @ a single Attribute

Plugging values into PDF for num attribs

this is a univariate PDF

① $$f(x) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|\Sigma_i|}} \exp\left\{-\frac{(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)}{2}\right\}$$

$\prod_{j=1}^{d} \sigma_{ij}^2$

$\sum_{j=1}^{d} \frac{(x_j - \mu_{ij})^2}{\sigma_{ij}^2}$

② Plugging gives us:

$$P(x|c_i) = \frac{1}{(\sqrt{2\pi})^d \sqrt{\prod_{j=1}^{d} \sigma_{ij}^2}} \exp\left\{-\sum_{j=1}^{d} \frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2}\right\}$$

2 remains as part of gaussian dist

③ use distributive to take Π out of Brackets

$$\prod_{j=1}^{d} \left(\frac{1}{\sqrt{2\pi}\,\sigma_{ij}} \exp\left\{-\frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2}\right\}\right)$$

④ $\prod_{j=1}^{d} P(x_j|c_i)$

final form is the product of the likelihoods

To summarise what we did here

the joint probability has been decomposed into a product of the probability along each dimension X_j

↳ caused by indep Assump

Naive Bayes uses the sample mean:

$\hat{\mu}_i = (\hat{\mu}_{i1}, \ldots, \hat{\mu}_{id})^T$

Diag Sample covariance Mat $= \hat{\Sigma}_i = \text{diag}(\hat{\sigma}_{i1}^2, \ldots, \hat{\sigma}_{id}^2)$

for each class c_i

Thus the complexity is only ever

~~$O(n^2 d)$~~ $O(nd)$

Pred:

$$\hat{y} = \text{argmax}\left\{ P(c_i) \prod^{d} F(x_j \mid \hat{\mu}_{ij}, \hat{\sigma}_{ij}^2) \right\}$$

Prior

↳ likelihood

Multivariate

Naive Bayes: Categoric Attrib

Indep assumption leads to a simplification of the joint prob mass function

Bayes: $P(x|c_i) = f(v|\theta) = f(X_1 = e_1 \dots X_j \text{ equal } c_i)$

recall the func is just $\frac{n_i(v)}{n_i}$

this can be rewritten as:

v = single not vect

$$P(x|c_i) = \prod_{j=1}^{d} P(x_j|c) = \prod_{j=1}^{d} f(X_j = e_{jr_j} | c_i)$$

this probmass function for X_j

As w/ bayes we can use pseudocount if the counts are zero:

$$\hat{f}(v_j|c_i) = \frac{n_i(v_j) + 1}{n_i + m_j}$$