ZAki

Bayes Classifier

BC directly uses Bayes theorem to predict the class for the new data instance

it estimates the posterior prob for each class $P(c_i|x)$ & chooses the class w/ the largest prob

$$\hat{y} = \text{argmax}\{P(c_i|x)\}$$

Bayes theorem allows us to invert the posterior probability in terms of the likelihood & prior probability

$$P(c_i|x) = \frac{P(x|c_i) \cdot P(c_i)}{P(x)}$$

$P(x|c_i)$ is the likelihood defined as the prob of obs x given c_i is the true class

$P(c_i)$ prior prob of c_i

$P(x)$ Prob of obs x from any class

$P(x)$ is fixed for a given point, so

$$\hat{y} = \text{argmax}\{P(x|c_i)P(c_i)\}$$

Predicted class depends on the likelihood of that ~~teh~~ class taking into account its prior prob

We need to estimate the likli & prior probs from the training set

ESTIMATING THE PRIOR probs

D_i Denote SubSet of D labelled c_i

$$D_i = \{X_j^T \mid X_j \text{ has class } y_j = c_j\}$$

Size $D = |D| = n$, Size of $D_i = |D_i| = n_i$

Prior Prob has be estimated as:

$$\hat{P}(C_i) = \frac{n_i}{n}$$

Estimating the Likelihood

To estimate the likelihood $P(x|c_i)$, we have to estimate the joint probability of x across all the d dimensions

That is to say, we have to estimate $P(x = (x_1, x_2, ..., x_d) | c_i)$

the prob of seeing a data instance with these parameters given the class of c_i

There is a difference in approach if the all the dimensions are numeric

The approach can be parametric or non-parametric. Here we explore parametric

To do this we assume each class c_i is normally distributed around some mean μ_i & a covariance matrix Σ_i

And estimated from the train set D.

for any class C_i, the prob density at x is thus given as

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

↑ Det of Conver mat

This is the probability density function

We use this because the probability of a precise point is 0. hence we need to look @ a range (PDF)

This gives use the relative values of likelihood of being near X

the PDF doesnt give us the probability directly. instead the prob density - prob per unit of X

to get the actual prob we need to consider an interval around X

the prob of x falling within that interval is given by the area under the pdf curve with ~~that~~ interval

When we calculate PDF, we are calcing so for class C_i @ point x

the part of the PDF that make this calc class specific is the inclusion of the classes μ_{MEAN} & Σ_{VAR}

recall the mean of a dataset will be a vector of the mean of each Dim

variance will = convariance matrix where Diags are indiv attribute variance & rest are coveriance relationship

We compute the likelihood by considering a asmall interval $\epsilon > 0$ centered @ x

$$P(X|C_i) = 2\epsilon \cdot f_i(x)$$

we use two because it represent two equal area both sides of X

ϵ is a tiny number

the posterior probability is then given:

Bayes

likelihood Prior

Proportion no 2ϵ

$$P(C_i|X) = \frac{2\epsilon \cdot f_i(x)\, P(C_i)}{\sum 2\epsilon \cdot f_i(x_j) P(C_j)} = \frac{f_i(x) P(C_i)}{\sum f_i(x_j) P(C_j)}$$

Same for all class

Note if we want to calc the likelihood we need the 2ϵ interval to get the area

However, in Bayes classifier we are interest in proportions between a class vs all classes

Hence the interval can be cancelled out

the bottom/Denominator is the Marginal Prob
~~Prob~~
The prob of observing X over all classes
- or regardless of class

This value will remain fix/constant
for all calculations

that is for calcing the prob that X
belongs to each class

Because it is constant we can ignore
~~Ingrad~~
So, the calculation becomes:

$$\hat{y} = \text{argmax} \{ f_i(x) P(c_i) \}$$

See that we just select the highest value = Pred

note, if we want to see the actual probs
and compare we would need the
full calc of the Denominator

Here, the 2ϵ still cancels out

only need to include 2ϵ if you want to scope even further in and get the likelihood

Calculating mean & variance Paras

Bayes uses sample mean & Sample covariant

$$\text{mean} = \hat{\mu}_i = \frac{1}{n_i} \sum_{x_j \in D_i} x_j$$

$$\text{Covar} = \hat{\Sigma} = \frac{1}{n_i} \bar{D}_i^T \bar{D}_i$$

where

$\bar{D}_i$ = centered matrix, Calc mean of col & subtract from all values

$$\bar{D}_i = D_i - 1 \cdot \hat{\mu}_i^T$$

these values are then plugged into the pdf eqs above

Bayes Categorical Attributes

Let X_j be a categorical attribute

It can take M_j distinct categorical values

$$Dom(X_j) = \{a_{j1}, a_{j2}, \ldots, a_{jM_j}\}$$

This is the set of values the attribute can take with a_{j1} being a particular value

Each categorical attribute is modelled as a M_j-Dimensional Multivariate Bernoulli random variable

Bernoulli = two categories

Multivariate = >2 categories rep as vector

M_j Dimensional = length of vector, num of cats

Vector values $e_{j1}, e_{j2}, \ldots, e_{jM}$ where e_{jr} is the r^{th} standard Basis vector in R^{M_j}

and corresponds to the r^{th} value
is $a_{jR} \in Dom(X_j)$ (the var categories)

Standard basis vector = $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$
Hot on encoding

where the 1 selects the category

$e_{jR} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = a_{jR} = a_{j3}$

why create this Multivariate Bernoulli Variable?

- it is not just about rewriting in a new format
- original Cat simply tells you what category and data point belongs to
- the random variable allows you to express probabilities with each variables
- Essential for handle uncertainty & prediction
- it also allows to conduct mathematical operations on categorical data & algos

the entire D-Dimensional dataset is Modelled as the vector RV = X

$$X = (X_1, X_2, \ldots, X_d)^T$$

This of this as tabular dataset

- X_n = the Column
- each row would be a data instand
- and each row holds a set of vectors which are the hot one encoded such

For the columns attributes

Let $d' = \sum_{j=1}^{d} M_j$

this is an important step in transforming the dataset into a format suitable for probabalistic modeling w/ categorical Attr

it is a way to represent the total Dimension of dataset after all Attriss have been hot one encoded

~~work w/ singe encoded vector~~

d' is the sum of the number of categories accross all attributes

i.e. total number of columns if every thing was not one encoded

it is also the length of a single concatenated vector of all attribs

Example

X_1 : Colour $\{red, green\}$ $M_1 = 2$

X_2 : Shape $\{Sq, Ci, tri\}$ $M_2 = 3$

X_2 : Size $\{sm, lg\}$ $M_3 = 2$

$d = 3$ = original Dimension

$d' = 7$ $M_1 + M_2 + M_3 = 2 + 3 + 2 = 7$

A categorical data point is represented

$$X = (X_1, X_2, \ldots, X_d)^T$$

where this is a data point where all Dimensions are catagorical

Therefore the d'-Dimensional Binary vector is:

Basically as hot one encoding

$$V = \begin{pmatrix} V_1 \\ \vdots \\ V_d \end{pmatrix} = \begin{pmatrix} e_{1r_1} \\ \vdots \\ e_{dr_d} \end{pmatrix}$$

$$V_j = e_{jr_j}$$

tells us the vector is the correp Standard basis Hot-one from within each Attrib j

tells the vector V is made up of other vector from Hot-one

The Probability of Point X (data point)

(PMF)

is obtained from the joint prob mass function for the vector Random variable X:

$$P(X|C_i) = f(v|C_i) = f(X_1 = e_{1_{f1}}, \ldots, X_D = e_{d_{rd}} \mid C_i)$$

$P(X|C_i)$ the prob of obs X data point given that it belongs to class C_i

$f(v)$ reps a function (Prob mass function) appled to a vector (single hot one encoded)

$P(X_1 = e \ldots e | C_1)$ represents attributes of a data point, its hot one-encoded vector (e) & again belonging to class C_i w/ the PMF function applied

the PMF can be directly estimated as

$$\hat{f}(v|C_i) \quad \frac{n_i(v)}{n_i}$$

$n_i(v)$ count ~~ther~~ exact vector occurs in n_i

Note that if the PMF (liklihood) comes out as 0, i.e. No vector counts for the class, then Bayes Classify will come out as 0 prob. This is an issue for a number of reasons:

1. indicates limited ~~data~~/missis data/sample
2. risks overfitting as model will be certain it won't occur
3. Violation of Bayes assump that data is indep - dependant data causes 0's
4. Math issue → x 0 = 0 = nullify Results
   → log(0) = undefined

A Solution = Pseudo-Count

Assume a +1 count for each value
i.e. to assume X occurs atleast once
1 + actual observed count of v in C_i

the creates an adjusted probcbcty mass at v given as:

$$\hat{f}(v|c_i) = \frac{n_i(v) + 1}{n_i + \Pi_{j=1}^{D} M_j}$$

Π = Product, i.e. multiplcation

Product of all numbers M_j from 1 to D
i.e. each Attribute D

~~Value~~ $M_1 = 2$, $M_2 = 3$, $M_3 = 2$ Unique
$= 2 \times 3 \times 2 = 12$ ⟵ Combinations

unique combos based on all categorical Attr

PMF needs to then be calculated for each class to extract a prediction of class

Challenges with Bayes Class

the main issue is lack of data to estimate the joint prob density(numberical) or mass funcher (catagorical)

for numberic we need to calc $O(d^2)$ covariances
- As Dims/Attribs increase this requires the estimation of to many paras
- d^2 = complexity grows expon w/ Dimensions increasing

for cats we ned to calc joint prob for all posible combinations of Attribs

To reduce these issues we can use a reduced set of paras in Practice known as Naive Bayes Classifier