Week 8

Naive Bayes algorithm

Recall:

1.

$$P(X = x, Y = y) = P(X = x|Y = y).\, P(Y = y) = P(Y = y|X = x).\, P(X = x)$$

2.

X and Y are independent iff
$P(X = x|Y = y) = P(X = x)$ for all $x \in R_x, y \in R_y$

that is
$P(X = x, Y = y) = P(X = x).\, P(Y = y)$

3. Bayes rule

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

Generative model based algorithms:

Algorithm 1:

Set up:

Data:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

- $y_i \in \{0, 1\}$ - Binary classification problem
- $x_i \in \{0, 1\}^d$ - d binary features

- Generative story: To come up with $P(x, y)$
- How we can model $P(x, y)$?

$$P(x, y) = P(y)P(x|y)$$

- $P(y) =$ distribution of labels. we need $P(y = 1)$.
  - can be estimated
- $P(x|y)$ = distribution of features given labels.
  - can be estimated

How many parameters?

- one for $P(y = 1) = p$

For a given label (say $y = 1$), how many possiblities for features?

- 2^d

That is 2^d examples are possible, need to estimate $2^d - 1$ parameters.

- Same for $y = 0$

Total parameters : $1 + 2(2^d - 1) = 2^{d+1} - 1$

Too many parameters

If we assume that the features are conditionally independent given the label, then for any $x = [f_1, f_2, \ldots f_d]$

$$p(x|y) = P(f_1|y).P(f_2|y)\ldots.P(f_d|y)$$

That is we just need estimates for $P(f_i|y)$.

Number of parameters to estimate:

- one for $P(y = 1) = p$
- for a given label (say $y = 1$),
  - d parameters $p_j^{y=1} = P(f_j = 1|y)$
- same for $y = 0$

Total parameters = $2d + 1$

- manageable

We assume the class conditional independent feature.

Naive Bayes algorithm:

- Naive: Class conditional independence
- Bayes: to find $P(y|x)$ using $p(y)$ and $p(x|y)$.

Now we have data $\{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$.

- distribution of y is bernoulli with unknowm parameter p.
- distribution of $x|y$ can be optained by distributions of $f_i|y$ which is again bernoulli with parameter $p_j^{y_i}$.

We have samples (data), how we can use this sample to estimate the parameters?

Use MLE

- $$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i$$
- $$\hat{p}_j^y = P(f_j = 1 | y_i = y) = \frac{\sum_{i=1}^{n} \mathbb{1}\left(f_j = 1, y_i = y\right)}{\sum_{i=1}^{n} \mathbb{1}\left(y_i = y\right)}$$

Prediction:

given $x_{\text{test}} \in \{0,1\}^d$, what will be y_{test}?

- Find $P(y_{\text{test}} = 0 | x_{\text{test}})$ and $P(y_{\text{test}} = 1 | x_{\text{test}})$
- If $P(y_{\text{test}} = 0 | x_{\text{test}}) > P(y_{\text{test}} = 1 | x_{\text{test}})$, predict 0
- predict 1 otherwise

How to find $P(y_{\text{test}} = 0 | x_{\text{test}})$ and $P(y_{\text{test}} = 1 | x_{\text{test}})$?

- use Bayes rule:

$$P(y_{\text{test}} = 0|x_{\text{test}}) = \frac{P(x_{\text{test}}|y_{\text{test}} = 0).\, P(y_{\text{test}} = 0)}{P(x_{\text{test}})}$$

$$P(y_{\text{test}} = 1|x_{\text{test}}) = \frac{P(x_{\text{test}}|y_{\text{test}} = 1).\, P(y_{\text{test}} = 1)}{P(x_{\text{test}})}$$

As we are comparing and denominator is same for both, ignore it!

$$
\begin{aligned}
P(y_{\text{test}} = 1 | x_{\text{test}}) &= P(x_{\text{test}} | y_{\text{test}} = 1) . P(y_{\text{test}} = 1) \\
&= \left(\prod_{i=1}^{d} (\hat{p}_i^1)^{f_i} (1 - \hat{p}_i^1)^{1 - f_i} \right) \hat{p}
\end{aligned}
$$

Example:

Consider the following dataset:

f_1	f_2	f_3	y
1	1	1	1
1	1	1	0
0	0	0	1
0	0	0	0
1	0	0	1
1	0	1	1
0	1	0	0

Pitfalls in naive Bayes:

What if j^{th} feature value is 1 for all the points??

- $\hat{p}_j^0 = ??$
- $\hat{p}_j^1 = ??$

What will be

$$P(y_{\text{test}} = 1 | x_{\text{test}}) \propto \left(\prod_{i=1}^{d} (\hat{p}_i^1)^{f_i} (1 - \hat{p}_i^1)^{1 - f_i} \right) \hat{p}$$

$$P(y_{\text{test}} = 0 | x_{\text{test}}) \propto \left(\prod_{i=1}^{d} (\hat{p}_i^0)^{f_i} (1 - \hat{p}_i^0)^{1 - f_i} \right) (1 - \hat{p})$$

What if j^{th} feature value is 0 for all the values??

- $\hat{p}_j^0 = ??$
- $\hat{p}_j^1 = ??$

What will be

$$P(y_{\text{test}} = 1 | x_{\text{test}}) = \left(\prod_{i=1}^{d} (\hat{p}_i^1)^{f_i} (1 - \hat{p}_i^1)^{1 - f_i} \right) \hat{p}$$

$$P(y_{\text{test}} = 0 | x_{\text{test}}) = \left(\prod_{i=1}^{d} (\hat{p}_i^0)^{f_i} (1 - \hat{p}_i^0)^{1 - f_i} \right) (1 - \hat{p})$$

Possible fix: Laplace smoothing

- Add $x = [1, 1, \ldots, 1]$ labeled as 0
- Add $x = [1, 1, \ldots, 1]$ labeled as 1
- Add $x = [0, 0, \ldots, 0]$ labeled as 0
- Add $x = [0, 0, \ldots, 0]$ labeled as 1

Decision function

Decision boundary is given by

$$\{x : P(y_{\text{test}} = 1|x_{\text{test}}) = P(y_{\text{test}} = 0|x_{\text{test}})\}$$

$$\left(\prod_{i=1}^{d}(\hat{p}_i^1)^{f_i}(1-\hat{p}_i^1)^{1-f_i}\right)\hat{p} = \left(\prod_{i=1}^{d}(\hat{p}_i^0)^{f_i}(1-\hat{p}_i^0)^{1-f_i}\right)(1-\hat{p})$$

Take $\log$ and manipulate, we get a linear function of features.

That is decision boundary in Bernoulli naive Bayes is linear.

Can we use Naive Bayes algorithm for regression problem?

Is Naive condition holds true in practice?

In practice, this assumption may not always be true, and there may be dependencies between the features that affect the accuracy of the Naive Bayes model.

- Naive Bayes has been found to be a useful and effective algorithm in many real-world applications, especially for text classification, spam filtering, and sentiment analysis.
- Naive Bayes is computationally efficient and can handle high-dimensional data with a large number of features, which makes it suitable for big data problems.

Gaussian naive Bayes: (for binary classification)

Assumption:

- naive condition: features given labels are conditionally independent.
- The features given labels follows the gaussian distribution.

$$x|y=0 \sim \text{Normal}(\mu_0, \Sigma_0)$$
$$x|y=1 \sim \text{Normal}(\mu_1, \Sigma_1)$$

$$\Sigma_0 = \Sigma_1$$

Parameters to estimate

- p
- μ_0, μ_1, Σ

- Use MLE
- $$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i$$
- $$\hat{\mu}_0 = \frac{\sum_{i=1}^{n} \mathbb{1}\left(y_i = 0\right) x_i}{\sum_{i=1}^{n} \mathbb{1}\left(y_i = 0\right)}$$
- $$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} \mathbb{1}\left(y_i = 1\right) x_i}{\sum_{i=1}^{n} \mathbb{1}\left(y_i = 1\right)}$$
- $$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu_{y_i})(x_i - \mu_{y_i})^T$$

Prediction

- Use Bayes rule
- Let f be the density of gaussian distribution
  - Predict $y_{\text{test}} = 1$ if

$$f(x_{\text{test}}, \hat{\mu}_1, \hat{\Sigma})\hat{p} > f(x_{\text{test}}, \hat{\mu}_0, \hat{\Sigma})(1 - \hat{p})$$

  - 0 otherwise

Decision function

- Linear if covariance matrices are the same.
- Quadratic if covariance matrices are different.