6. Naive Bayes classifier

Shabana K M
PhD Research Scholar
Computer Science and Engineering

IIT Palakkad

26 September 2021

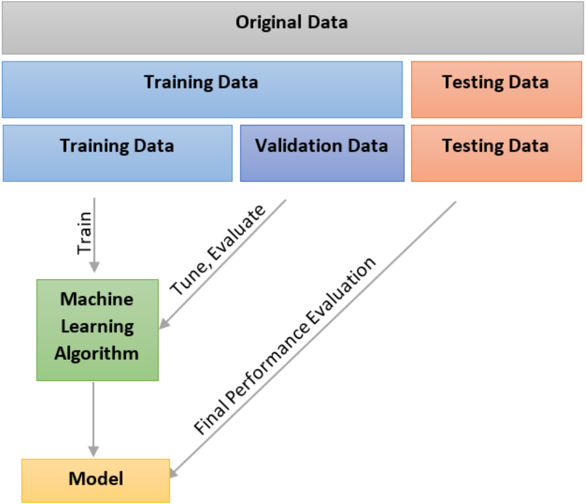
INDIAN INSTITUTE OF TECHNOLOGY PALAKKAD

IEEE STUDENT BRANCH GEC PALAKKAD IEEE LINK INTERNZONE IEEE Kerala Section IEEE

Recap

- Logistic regression
  - Gradient descent update
  - Multi-class classification - One vs all approach
- k-nearest neighbor algorithm
  - regression, classification
  - pros and cons
- Cross-validation
  - holdout method
  - k-fold
  - leave one out
  - stratified k-fold

Cross validation: The idea



Cross validation techniques

Figure: Holdout method

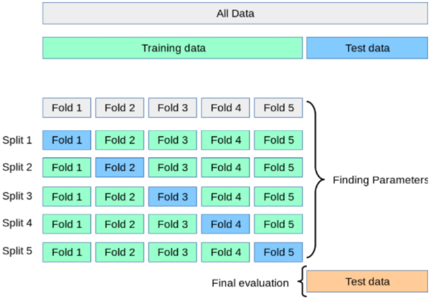
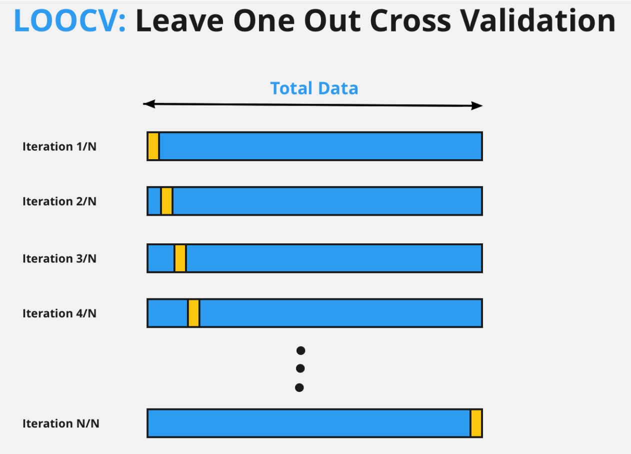


Figure: k-fold cross validation

Cross validation techniques



Cross validation techniques

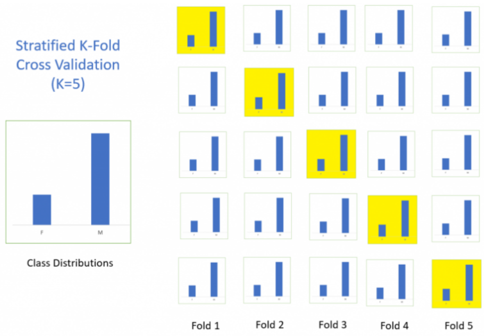
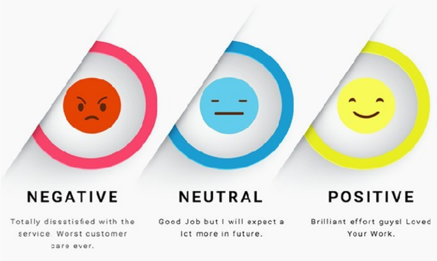


Figure: stratified k-fold cross validation

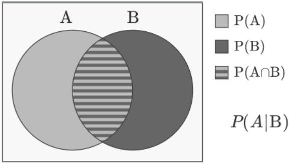
Motivation


Raw Text Data
Technology
Sports
Fashion


SPAM
SPAM FILTER


NEGATIVE
Totally dissatisfied with the service. Worst customer care ever.
NEUTRAL
Good Job but I will expect a lot more in future.
POSITIVE
Brilliant effort guys! Loved Your Work.

Conditional probability and Bayes theorem



$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Probability of B occurring given evidence A has already occurred

Probability of A occurring

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Probability of A occurring given evidence B has already occurred

Probability of B occurring

Bayes theorem - example

Given

- A doctor knows that Cold causes fever 50% of the time
- Prior probability of any patient having cold is 1/50,000
- Prior probability of any patient having fever is 1/20

If a patient has fever, what's the probability he/she has cold?

$$P(cold|fever) = \frac{P(fever|cold) * P(cold)}{P(fever)}$$

$$P(cold|fever) = \frac{0.5 * 1/50000}{1/20} = 0.0002$$

Naive Bayes classifier

- supervised machine learning algorithm based on Bayes theorem
- makes strong independence assumption on features
  - assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature
  - for instance, in gender prediction, a person's height and weight is assumed to independently contribute to the probability that a person is male/female
  - naive assumption
- demonstrated good performance in many complex real-world applications
- easy to build and particularly useful for very large data sets
- only requires a small amount of training data to estimate parameters

Bayesian classifiers

- features and class variable considered as random variables
- given a data point with attributes $(x_1, x_2, ..., x_D)$
  - goal is to predict the target class y
  - specifically, find the value of y that maximizes $P(y|x_1, x_2, ..., x_D)$
- how to estimate $P(y|x_1, x_2, ..., x_D)$ directly from data?

Bayesian classifiers

Approach: Use Bayes theorem

$$P(y|x_1, x_2, ..., x_D) = \frac{P(x_1, x_2, ..., x_D|y)P(y)}{P(x_1, x_2, ..., x_D)}$$

- choose value of y that maximizes $P(y|x_1, x_2, ..., x_D)$
- equivalent to choosing value of y that maximizes $P(x_1, x_2, ..., x_D|y)P(y)$
  - since the denominator doesn't depend on the class y, it can be considered a constant

How to estimate $P(x_1, x_2, ..., x_D|y)$ and $P(y)$ **?**

Naive Bayes classifier

Assume all features are `independent` when class is given

$$P(x_1, x_2, ..., x_D|y = i) = P(x_1|y = i)P(x_2|y = i)...P(x_D|y = i)$$

$P(x_j|y = i)$ can be estimated for all values of $j = 1, ..., D$ and $i = 1, ..., L$

- D: number of features
- L: number of classes

A new data point is classified to class k if $P(y = k) \prod_{j=1}^{D} P(x_j|y = k)$ is maximum over all values of y

- $\prod_{j=1}^{D} P(x_j|y = k) = P(x_1|y = k)P(x_2|y = k)...P(x_D|y = k)$

Estimating class probabilities from data

- **Class probabilities:** $P(y = k)$

$$P(y = k) = \frac{\text{Number of observations with } y = k}{\text{Total number of observations}}$$

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

- $\mathrm{P(y = Yes) = 3/10}$
- $\mathrm{P(y = No) = 7/10}$

Estimating conditional probabilities - Discrete attributes

- **Conditional probabilities:** $P(x_j|y = k)$

$$P(x_j = z|y = k) = \frac{\text{No: of observations with class } y = k \text{ having } x_j = z}{\text{Total number of observations with class } y = k}$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- $P(Status = Married|y = No) = 4/7$
- $P(Status = Single|y = No) = 2/7$
- $P(Status = Divorced|y = No) = 1/7$
- $P(Status = Married|y = Yes) = 0$

Sample correction

- if a given class and feature value never occur together in the training set then the frequency-based probability estimate will be zero

$$\text{if } P(x_i|y=k)=0, \text{ for some } 1 \leq i \leq D, \text{ then}$$
$$P(x_1|y=k)P(x_2|y=k)...P(x_D|y=k)=0$$

- this is undesirable as the information in the other attributes is lost
- a small sample correction or pseudo-count is employed such that no probability is ever set to be exactly zero - `smoothing techniques`

Estimating conditional probabilities - Continuous attributes

- **Discretize** the range into bins
  - define bins for the attribute
  - use one ordinal attribute per bin
  - violates independence assumption
- **Probability density** estimation
  - assume that the attribute follows a normal distribution
  - use data to estimate parameters of the distribution (mean and standard deviation)
  - estimate the conditional probability $P(x_i|y = j)$ based on this distribution

Naive Bayes - Example

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Naive Bayes - Example

Learning Phase

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=*Yes*	Play=No
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=*Yes*	Play=*No*
Strong	3/9	3/5
Weak	6/9	2/5

$P(\text{Play}=\textit{Yes}) = 9/14 \qquad P(\text{Play}=\textit{No}) = 5/14$

Naive Bayes - Example

Given a new observation $x' =$ (Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong), predict its label (`Play = Yes/No`)

- **Compute conditional probabilities**

P(Outlook=*Sunny* | Play=*Yes*) = 2/9
P(Temperature=*Cool* | Play=*Yes*) = 3/9
P(Huminity=*High* | Play=*Yes*) = 3/9
P(Wind=*Strong* | Play=*Yes*) = 3/9
P(Play=*Yes*) = 9/14

P(Outlook=*Sunny* | Play=*No*) = 3/5
P(Temperature=*Cool* | Play==*No*) = 1/5
P(Huminity=*High* | Play=*No*) = 4/5
P(Wind=*Strong* | Play=*No*) = 3/5
P(Play=*No*) = 5/14

- **Compute $P(y|x')$ using Bayes rule**

$$P(Yes \mid \mathbf{x}') \approx [P(Sunny \mid Yes)P(Cool \mid Yes)P(High \mid Yes)P(Strong \mid Yes)]P(Play{=}Yes) = 0.0053$$
$$P(No \mid \mathbf{x}') \approx [P(Sunny \mid No)\, P(Cool \mid No)P(High \mid No)P(Strong \mid No)]P(Play{=}No) = 0.0206$$

Since $P(Yes|x') < P(No|x')$, x' is labeled with **Play = No**

Bag of words model

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

Spam detection

	Label	SMS
0	spam	SECRET PRIZE! CLAIM SECRET PRIZE NOW!!
1	ham	Coming to my secret party?
2	spam	Winner! Claim secret prize now!

$$\downarrow$$

Label	secret	prize	claim	now	coming	to	my	party	winner
spam	2	2	1	1	0	0	0	0	0
ham	1	0	0	0	1	1	1	1	0
spam	1	1	1	1	0	0	0	0	1

$$P(\mathrm{Spam}|w_1, w_2, \ldots, w_n) \propto P(\mathrm{Spam}) \cdot \prod_{i=1}^{n} P(w_i|\mathrm{Spam})$$

$$P(\mathrm{Ham}|w_1, w_2, \ldots, w_n) \propto P(\mathrm{Ham}) \cdot \prod_{i=1}^{n} P(w_i|\mathrm{Ham})$$

References

1 https://web.iitd.ac.in/~bspanda/BY.pdf

Thanks Google for the pictures!