Machine Learning

Week 8
Naïve Bayes Classification

Dr. Nouman M Durrani

Acknowledgement to all authors whose materials have been used

Naive Bayes Classification

- Suppose you are a product manager, you want to classify:
  - Customer reviews in positive and negative classes.
  - As a loan manager, you want to identify which loan applicants are safe or risky?
  - As a healthcare analyst, you want to predict which patients can suffer from diabetes disease.
- All the examples have the same kind of problem to classify reviews, loan applicants, and patients.
- Naive Bayes is the most straightforward and fast classification algorithm, which is suitable for a large chunk of data.
- Naive Bayes classifier is successfully used in various applications such as:
  - spam filtering, text classification, sentiment analysis, and recommender systems.
- It uses Bayes theorem of probability for prediction of unknown class.

Naive Bayes Classifier

- A Naive Bayes classifier is a probabilistic machine learning model that is used for classification task. The classifier is based on the Bayes theorem.

- We can find the probability of **A** happening, given that **B** has occurred by the Bayes Theorem as:

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

- Here, **B** is the evidence and **A** is the hypothesis.

- P(A\B) : Probability of A when B is True
- P(A) : Probability of Hypothesis
- P(B): Probability of Evidence
- P(B\A) : Probability of B when A is True

Bayes Theorem

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- P(h): the probability of hypothesis h being true (regardless of the data). This is known as the prior probability of h.
- P(D): the probability of the data (regardless of the hypothesis). This is known as the prior probability.
- P(h|D): the probability of hypothesis h given the data D. This is known as posterior probability.
- P(D|h): the probability of data d given that the hypothesis h was true. This is known as posterior probability.

Naive Bayes Classifier

- The Naive Bayes classifier assumes that the predictors/features are independent.
  - The presence of one particular feature does not affect the other.
  - The presence of a feature in a class is unrelated to any other feature.
- Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that a particular fruit is an apple or an orange or a banana, and that is why it is known as **"Naive."**
- Naive Bayes classifiers have high accuracy and speed on large datasets.

Consider the problem of playing golf. The dataset is represented as below.

We classify whether the day is suitable for playing golf, given the features of the day. According to this example, Bayes theorem can be written as:

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

The variable **y** is the class variable(play golf), which represents if it is suitable to play golf or not given the conditions X. Variable **X** represent the parameters/features. **X** is given as,

$$X = (x_1, x_2, x_3, ..., x_n)$$

$$P(y|x_1, ..., x_n) =$$

$$\frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY GOLF
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	Mild	High	False	No
8	Rainy	Cool	Normal	False	Yes
9	Sunny	Mild	Normal	False	Yes
10	Rainy	Mild	Normal	True	Yes
11	Overcast	Mild	High	True	Yes
12	Overcast	Hot	Normal	False	Yes
13	Sunny	Mild	High	True	No

HOW NAIVE BAYES CLASSIFIER WORKS?

Whether	Temperature	Play
Sunny	Hot	No
Sunny	Hot	No
Overcast	Hot	Yes
Rainy	Mild	Yes
Rainy	Cool	Yes
Rainy	Cool	No
Overcast	Cool	Yes
Sunny	Mild	No
Sunny	Cool	Yes
Rainy	Mild	Yes
Sunny	Mild	Yes
Overcast	Mild	Yes
Overcast	Hot	Yes
Rainy	Mild	No

01 CALCULATE PRIOR PROBABILITY FOR GIVEN CLASS LABELS

02 CALCULATE CONDITIONAL PROBABILITY WITH EACH ATTRIBUTE FOR EACH CLASS

03 MULTIPLY SAME CLASS CONDITIONAL PROBABILITY.

04 MULTIPLY PRIOR PROBABILITY WITH STEP 3 PROBABILITY.

05 SEE WHICH CLASS HAS HIGHER PROBABILITY, HIGHER PROBABILITY CLASS BELONGS TO GIVEN INPUT SET STEP.

Question on NAIVE BAYE's Algorithm:

Ques:) For the given dataset, Apply Naive-Baye's algorithm and Predict the outcome for a Car = { Red, Domestic, SUV }

Color	Type	Origin	Stolen
Red	Sports	Domestic	Yes
Red	Sports	Domestic	NO
Red	Sports	Domestic	Yes
Yellow	Sports	Domestic	NO
Yellow	Sports	Imported	Yes
Yellow	SUV	Imported	NO
Yellow	SUV	Imported	Yes
Yellow	SUV	Domestic	NO
Red	SUV	Imported	NO
Red	Sports	Imported	Yes

Posterior $P(A|B) = \frac{P(B|A).P(A)}{P(B)}$ → Proposition

Prob. of B when A is True (Likelihood)

Prob (A) when B is true

evidence

NO ANS..

$X = [\text{Red}, \text{Domestic}, \text{SUV}] = P(X|\text{Yes}) = \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} = \frac{6}{125} = 0.024$ Yes

i) $P(\text{Red}|\text{Yes}) = \frac{P(\text{Yes}|\text{Red}).P(\text{Red})}{P(\text{Yes})} = \frac{\frac{3}{5} \cdot \frac{5}{10} \cdot \frac{3}{5}}{5/10}$

ii) $P(\text{Domestic}|\text{Yes}) = 2/5$ iii) $P(\text{SUV}|\text{Yes}) = 1/5$

$P(\text{Red}|\text{No}) = \frac{P(\text{No}|\text{Red}).P(\text{Red})}{P(\text{No})} = \frac{\frac{2}{5} \cdot \frac{5}{10}}{5/10} = 2/5$

NO

$P(\text{Domestic}|\text{No}) = 3/5$, $P(\text{SUV}|\text{No}) = 2/5$ $= \frac{2}{5} \times \frac{3}{5} \times \frac{2}{5} = 0.072$

Bayes' Theorem Example

- Let's suppose we have a Deck of Cards and we wish to find out the probability of the card we picked at random to being a king, given that it is a face card. So, according to Bayes' Theorem, we can solve this problem. First, we need to find out the probability:
- **P(King)** which is **4/52** as there are 4 Kings in a Deck of Cards.
- **P(Face|King)** is equal to **1** as all the Kings are face Cards.
- **P(Face)** is equal to **12/52** as there are 3 Face Cards in a Suit of 13 cards and there are 4 Suits in total.

$$P(King|Face) = \frac{P(Face|King).P(King)}{P(Face)}$$

$$= \frac{1.(1/13)}{3/13} = 1/3$$

$$P(King) = 4/52 = 1/13$$

$$P(Face|King) = 1$$

$$P(Face) = 12/52 = 3/13$$

Naïve Bayes Classifier

Example:

- *Given Outlook, Temperature, Humidity and Wind Information, we want to carry out prediction for Play: Yes or No.*

- Mathematically, which one is greater

$$P(\text{Play} = \text{Yes} \mid \text{Outlook},\ \text{Temp.},\ \text{Humidity},\ \text{Wind})$$

$$P(\text{Play} = \text{No} \mid \text{Outlook},\ \text{Temp.},\ \text{Humidity},\ \text{Wind})$$

- Predict for Sunny outlook, High humidity, Cool temperatue and Weak wind.

- Predict the most likely.

Day	Outlook	Temp.	Humidity	Wind	Play
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

Reference: Section 6.9.1 (Machine Learning by Tom Mitchell)

Naïve Bayes Classifier

Example:

$$P(\text{Play} = \text{Yes} \mid \text{Outlook} = \text{Sunny},\ \text{Temp} = \text{Cool},\ \text{Humidity} = \text{High},\ \text{Wind} = \text{Weak})$$

$$= \frac{P(\text{Outlook} = \text{Sunny},\ \text{Temp} = \text{Cool},\ \text{Humidity} = \text{High},\ \text{Wind} = \text{Strong} \mid \text{Play} = \text{Yes})\, P(\text{Play} = \text{Yes})}{P(\text{Outlook} = \text{Sunny},\ \text{Temp} = \text{Cool},\ \text{Humidity} = \text{High},\ \text{Wind} = \text{Strong})}$$

Naïve Assumption:

- Feature are mutually independent given the label!

$$P(\text{Outlook} = \text{Sunny},\ \text{Temp} = \text{Cool},\ \text{Humidity} = \text{High},\ \text{Wind} = \text{Strong} \mid \text{Play} = \text{Yes})$$

$$= P(\text{Outlook} = \text{Sunny} \mid \text{Play} = \text{Yes})\, P(\text{Temp} = \text{Cool} \mid \text{Play} = \text{Yes})\, P(\text{Humidity} = \text{High} \mid \text{Play} = \text{Yes})\, P(\text{Wind} = \text{Strong} \mid \text{Play} = \text{Yes})$$

Naïve Bayes Classifier

Example:

$$P(\text{Outlook} = \text{Sunny} \mid \text{Play} = \text{Yes}) = \frac{2}{9}$$

$$P(\text{Temp} = \text{Cool} \mid \text{Play} = \text{Yes}) = \frac{3}{9}$$

$$P(\text{Humidity} = \text{High} \mid \text{Play} = \text{Yes}) = \frac{3}{9}$$

$$P(\text{Wind} = \text{Strong} \mid \text{Play} = \text{Yes}) = \frac{3}{9}$$

$$P(\text{Play} = \text{Yes}) = \frac{9}{14}$$

$$P(\text{Outlook} = \text{Sunny} \mid \text{Play} = \text{No}) = \frac{3}{5}$$

$$P(\text{Temp} = \text{Cool} \mid \text{Play} = \text{No}) = \frac{1}{5}$$

$$P(\text{Humidity} = \text{High} \mid \text{Play} = \text{No}) = \frac{4}{5}$$

$$P(\text{Wind} = \text{Strong} \mid \text{Play} = \text{No}) = \frac{3}{5}$$

$$P(\text{Play} = \text{No}) = \frac{5}{14}$$

Day	Outlook	Temp.	Humidity	Wind	Play
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

Naïve Bayes Classifier

Example:

$$P(\text{Outlook} = \text{Sunny} \mid \text{Play} = \text{Yes})\ P(\text{Temp} = \text{Cool} \mid \text{Play} = \text{Yes})\ P(\text{Humidity} = \text{High} \mid \text{Play} = \text{Yes})\ P(\text{Wind} = \text{Strong} \mid \text{Play} = \text{Yes})$$

$$\times P(\text{Play} = \text{Yes}) \qquad = \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14} \qquad = 0.0053$$

$$P(\text{Outlook} = \text{Sunny} \mid \text{Play} = \text{No})\ P(\text{Temp} = \text{Cool} \mid \text{Play} = \text{No})\ P(\text{Humidity} = \text{High} \mid \text{Play} = \text{No})\ P(\text{Wind} = \text{Strong} \mid \text{Play} = \text{No})$$

$$\times P(\text{Play} = \text{No}) \qquad = \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14} \qquad = 0.0206$$

$$P(\text{Play} = \text{Yes} \mid \text{Outlook} = \text{Sunny},\ \text{Temp} = \text{Cool},\ \text{Humidity} = \text{High},\ \text{Wind} = \text{Strong}) = \frac{0.0053}{0.0053 + 0.0206} = 0.2046$$

$$P(\text{Play} = \text{No} \mid \text{Outlook} = \text{Sunny},\ \text{Temp} = \text{Cool},\ \text{Humidity} = \text{High},\ \text{Wind} = \text{Strong}) = \frac{0.0206}{53 + 0.0206} = 0.7954$$

Play = No *is more likely!*

Naïve Bayes Classifier

Generative Classifier:

- Attempts to model class, that is, build a generative statistical model that informs us how a given class would generate input data.

- Ideally, we want to learn the joint distribution of the input **x** and output label y, that is, P(**x**,y).

- For a test-point, generative classifiers predict which class would have **most-likely** generated the given observation.

- Mathematically, prediction for input **x** is carried out by computing the conditional probability P(y|**x**) and selecting the most-likely label y.

- Using the Bayes rule, we can compute P(y|**x**) by computing P(y) and P(**x**|y).

- Estimating $P(y)$ and $P(\mathbf{x}|y)$ is called generative learning.

Naïve Bayes Classifier

Overview of Naïve Bayes Classifier:

- We have $D = \{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \ldots, (\mathbf{x_n}, y_n)\} \subseteq \mathcal{X}^d \times \mathcal{Y}$

  $\mathcal{Y} = \{1, 2, \ldots, M\}$ (M-class classification)

Key Idea:

- Estimate $P(y|\mathbf{x})$ from the data using the Bayes Theorem.
- Using Bayes theorem and MAP learning framework, we can write this as

$$h_{\text{MAP}}(\mathbf{x}) = \underset{y \in \mathcal{Y}}{\text{maximize}} \quad P(y \mid \mathbf{x}) = \underset{y \in \mathcal{Y}}{\text{maximize}} \quad \frac{P(\mathbf{x} \mid y)\, P(y)}{P(\mathbf{x})} = \underset{y \in \mathcal{Y}}{\text{maximize}} \quad P(\mathbf{x} \mid y)\, P(y)$$

- Estimating $P(y)$ is easy. If y takes on discrete binary values, coin tossing or spam vs non-spam for example, we simply need to count how many times we observe each class outcome.
- Estimating $P(\mathbf{x}|y)$, however, is not easy, Why?

Naïve Bayes Classifier

Overview of Naïve Bayes Classifier:

Example:

- $M = 2$ and features $d = 6$. Assuming binary features/classification.

- We want to estimate

$$P(\mathbf{x} \mid y) = P(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, x^{(5)}, x^{(6)} | y)$$

- How many parameters do we need to fully estimate $P(\mathbf{x}|y)$?

- We need to represent all 2^6 outcomes or probabilities for each $y = 0, 1$.

- For d binary features, we need to represent all 2^d outcomes.

- Learning the values for the full conditional probability would require enormous amounts of data.

time	Inputv1	Inputv2	Inputv3	Inputv4	Inputv5	Inputv6	output
19:50:00	1	0	0	1	0	0	1
19:55:00	1	0	0	1	0	0	0
20:00:00	1	0	0	1	0	0	1
20:05:00	1	1	1	0	0	0	1
20:10:00	1	1	1	0	0	0	1
20:15:00	1	1	0	1	0	0	1
20:20:00	1	1	0	1	1	0	0
20:25:00	1	0	0	1	1	0	1
20:30:00	1	0	0	1	1	0	1
20:35:00	0	0	0	1	0	0	1
20:40:00	1	0	0	1	1	0	1
20:45:00	0	0	0	1	0	0	0

Naïve Bayes Classifier

Naïve Bayes Classifier:

- To overcome this requirement of enormous data for the computation of conditional probability, we can make a 'naive Bayes' assumption.

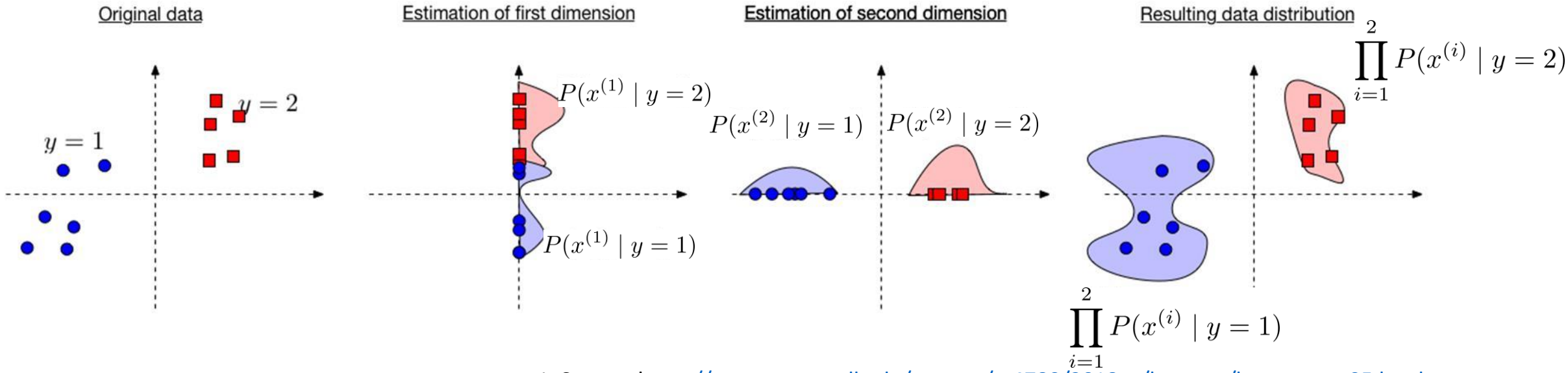
Naïve Assumption:

- Features are mutually independent given the label!
- Consequence: $P(\mathbf{x} \mid y) = P(x^{(1)}, x^{(2)}, \ldots, x^{(d)}|y) = \prod_{i=1}^{d} P(x^{(i)} \mid y)$
- How many probabilities now?

  one for each feature/label.

  $2d$

Interpretation[1]:



1. Source: https://www.cs.cornell.edu/courses/cs4780/2018sp/lectures/lecturenote05.html

Naïve Bayes Classifier

Naïve Bayes Classifier:

- We can reformulate our hypothesis function, referred to as Naive Bayes (NB) Classifier, as

$$h_{\text{NB}}(\mathbf{x}) = \underset{y \in \mathcal{Y}}{\text{maximize}} \quad P(y \mid \mathbf{x}) = \underset{y \in \mathcal{Y}}{\text{maximize}} \quad \prod_{i=1}^{d} P(x^{(i)} \mid y)\, P(y)$$

- Maximizes the log (natural, ln) of the function instead.

$$h_{\text{NB}}(\mathbf{x}) = \underset{y \in \mathcal{Y}}{\text{maximize}} \quad \sum_{i=1}^{d} \log \left(P(x^{(i)} \mid y)\, P(y) \right)$$

$$= \underset{y \in \mathcal{Y}}{\text{maximize}} \quad \sum_{i=1}^{d} \log P(x^{(i)} \mid y) \; + \; \log P(y)$$

- How many probabilities?

$2d + 1$

Naïve Bayes Classifier

Naïve Bayes Classifier - Training:

Assume each feature and label as a binary variable

- Hypothesis space: $2d + 1$ different binomial distributions.
  - $P(x^{(i)} \mid y)$ and $P(y)$ for each $x^{(i)}$ and each $y = \{0, 1\}$, $i = 1, 2, \ldots, d$.
  - Each probability can be parameteterized by a single variable θ.
- We treat learning of each of these as a separate MLE problem.

$$P(x^{(i)} = j \mid y = k) = \frac{\text{count}(x^{(i)} = j \text{ and } y = k)}{\text{count}(y = k)}, \quad j, k \in \{0, 1\}$$

$$P(y = k) = \frac{\text{count}(y = k)}{\text{count}(y = 0) + \text{count}(y = 1)} = \frac{\text{count}(y = k)}{n}, \quad k \in \{0, 1\}$$

- We compute these probabilities during training stage.
- As we saw earlier, these probability estimates maximizes the likelihood.

Naïve Bayes Classifier

Naïve Bayes Classifier - Prediction:

Assume each feature and label as a binary variable

- For a new test-point $\mathbf{x}_{\text{new}}$, we assign the label as

$$h_{\text{NB}}(\mathbf{x}_{\text{new}}) = \underset{y \in \mathcal{Y}}{\text{maximize}} \quad P(y \mid \mathbf{x}_{\text{new}}) = \underset{y \in \mathcal{Y}}{\text{maximize}} \quad \prod_{i=1}^{d} P(x_{\text{new}}^{(i)} \mid y)\, P(y)$$

We have a problem here!

- We have a product of probabilities. If any of the estimated probability is zero, the product would be zero.

Solution: Additive Smoothing or Laplace Smoothing

$$P(x^{(i)} = j \mid y = k) = \frac{\text{count}(x^{(i)} = j \text{ and } y = k) + \ell}{\text{count}(y = k) + \ell R}, \quad j, k \in \{0, 1\}$$

$$P(y = k) = \frac{\text{count}(y = k) + \ell}{n + \ell M}, \quad k \in \{0, 1\}$$

- Here $\ell > 0$. If $\ell = 1$, we refer to it as add-1 smoothing.
- R is the number of values $x^{(i)}$ can take. For binary case, $R = 2$.
- M is the number of classes. For binary case $M = 2$.

Naïve Bayes Classifier

Naïve Bayes Classifier - Extensions:

- We have $D = \{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \ldots, (\mathbf{x_n}, y_n)\} \subseteq \mathcal{X}^d \times \mathcal{Y}$

  $\mathcal{Y} = \{1, 2, \ldots, M\}$ (M-class classification)

- We assume that each feature $x^{(i)}$ takes L_i values, that is, $x^{(i)} \in \{1, 2, \ldots, L_i\}$.

How many probability tables do we have if we have d features and M labels?

- $dM + 1$: we have one probability table for each feature and each value of the label and one more table for the prior $P(y)$.
- The set of tables for a single feature (for all labels y) is referred to as a conditional probability table (CPT), and here we have d of those.

Incorporating model parameters in the formulation

- We considered a binary case and assumed that a single parameter characterizes probability model associated with each feature.
- In general, we can have parameters defining the probability model and we learn parameters of the probability model during the learning stage.

Naïve Bayes Classifier

Naïve Bayes Classifier - Summary:

- In Naïve Bayes, we compute the probabilities or parameters of the distribution defining probabilities and use these to carry out predictions.
- Naïve Bayes can handle missing values by ignoring the sample during probability computation, is robust to outliers and irrelevant features.
- Naïve Bayes algorithm is very easy to implement for applications involving textual information data (e.g., sentiment analysis, news article classification, spam filtering).
- Convergence is quicker relative to logistic regression (to be studied later) that discriminative in nature.
- It performs well even when the independence between features assumption does not hold.
- The resulting decision boundaries can be non-linear and/or piecewise.
- Disadvantage: It is not robust to redundant features. If the features have a strong relationship or correlation with each other, Naïve Bayes is not a good choice. Naïve Bayes has high bais and low variance and there are no regularization here to adjust the bias thing

NB Classifier – Text Classification

Text Classification Overview:

- Applications of text classification include
  - Sentiment analysis
  - Spam detection
  - Language Identification; to name a few.

Classification Problem:

Input: a document and a fixed set of classes (e.g., spam, non-spam)

Output: a predicted class for the document

Classification Methods:

- **Hand-coded rules:** Rules based on combinations of words or other features
  - e.g., spam: black-list-address OR ("dollars" AND "you have been selected")
  - Accuracy can be high if rules carefully refined by **expert**
  - But building and **maintaining** these rules is **expensive**

Ref: **"Speech and Language Processing"** by *Daniel Jurafsky & James H. Martin*
Source: https://web.stanford.edu/~jurafsky/slp3/

NB Classifier – Text Classification

Text Classification – Supervised Learning:

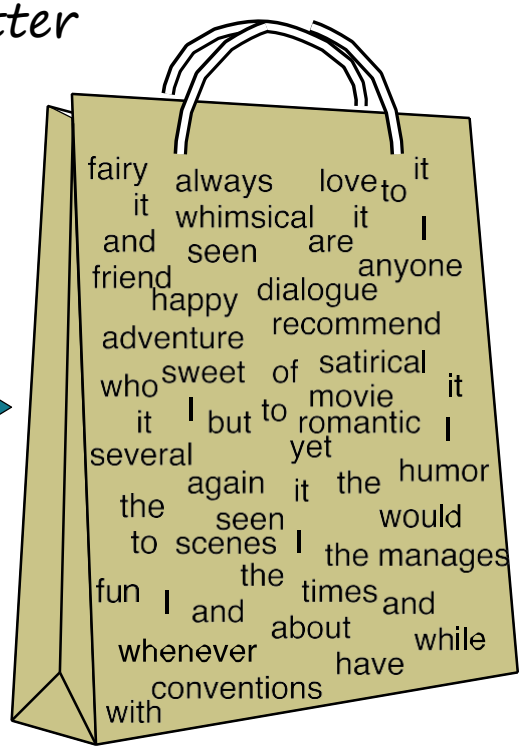
Input: *a document and a fixed set of classes (e.g., spam, non-spam)*
+ training data (n labeled documents)

Output: *a predicted class for the document*

Bag of Words – Representation of a document for classification:

Assumption: *Position doesn't matter*

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

Ref: "Speech and Language Processing" by *Daniel Jurafsky & James H. Martin*
Source: https://web.stanford.edu/~jurafsky/slp3/

NB Classifier – Text Classification

Text Classification – Terminology and Preprocessing :

- Corpus: A collection of documents; data.

- Vocabulary, denoted by V, is the union of all the word types in all classes (not just one class).

Preprocessing documents:

- Clean the corpus: (e.g., Hello, hello or hello! should be considered the same)
  - Remove numbers, punctuation and excessive white spaces
  - Use lowercase representation

- Stop words concept: very frequent words (**a** or **the**)
  - Sort vocabulary with respect to frequency, call the top 5 or 20 words the stopword list and remove from all of the documents or from the vocabulary.

- In naïve Bayes, it's more common to **not** remove stop words and use all the words.

- After pre-processing, create a **mega document** for each class by concatenating all the documents of the class.

- Use bag of words on mega document to obtain a frequency table for each class.

Ref: **"Speech and Language Processing"** by *Daniel Jurafsky & James H. Martin*
Source: https://web.stanford.edu/~jurafsky/slp3/

NB Classifier – Spam Filtering

Example: Spam vs Non-Spam:

Category	Document
Spam	send us your password
Spam	review us
Spam	send us your account
Spam	send your password
Non-spam	password review
Non-spam	send us your review
?	review us now
?	review account

- Vocabulary, V = {send, us, your, password, review, account}

- Issue 1:
- 'now' *is not in the training data.*
- *– unknown word or out of vocabulary word.*
- Solution:
- *remove out of vocabulary word from the test document.*

- Issue 2:
- 'account' *is only available in one class*
- Solution:

NB Classifier – Spam Filtering

Naïve Bayes (NB) Classification:

- NB Classifier:

$$h_{\text{NB}}(\mathbf{x}) = \underset{y \in \mathcal{Y}}{\text{maximize}} \quad P(y \mid \mathbf{x}) = \underset{y \in \mathcal{Y}}{\text{maximize}} \quad \prod_{i=1}^{d} P(x^{(i)} \mid y)\, P(y)$$

- $\mathbf{x}$ represents the test document for which we want to carry out prediction. Each feature represents a word in the document.
- d here represents the number of words in the test document.

- For $\mathbf{x}$ = "review us now", $d = 3$.

- For $\mathbf{x}$ = "review account", $d = 2$.

Ref: "Speech and Language Processing" by *Daniel Jurafsky & James H. Martin*
Source: https://web.stanford.edu/~jurafsky/slp3/

NB Classifier – Spam Filtering

Naïve Bayes (NB) Classification – Example:

Category	Document
Spam	send us your password
Spam	review us
Spam	send us your account
Spam	send your password
Non-spam	password review
Non-spam	send us your review
?	review us now
?	review account

Bag of Words

Vocabulary	Spam Count	Non-spam Count
send	3	1
us	3	1
your	3	1
password	2	1
review	1	2
account	1	0
	13	6

- For $\mathbf{x} =$"review us now", $d = 3$.

We compute $P(\text{Spam} \mid \mathbf{x})$ and $P(\text{Non} - \text{spam} \mid \mathbf{x})$

NB Classifier – Spam Filtering

Naïve Bayes (NB) Classification – Example:

- For $\mathbf{x}$ = "review us now".
- Ignore 'now': unknown word, out of vocabulary
- We compute $P(\mathbf{x} \mid \text{Spam})\, P(\text{Spam})$ and $P(\mathbf{x} \mid \text{Non}-\text{spam})\, P(\text{Non}-\text{spam})$

$$P(\mathbf{x} \mid \text{Spam})\, P(\text{Spam}) = P(\text{review} \mid \text{Spam}) P(\text{us} \mid \text{Spam})\, P(\text{Spam})$$

$$P(\text{review} \mid \text{Spam}) = \frac{1}{13} \qquad P(\text{us} \mid \text{Spam}) = \frac{3}{13} \qquad P(\text{Spam}) = \frac{4}{6}$$

$$P(\mathbf{x} \mid \text{Spam})\, P(\text{Spam}) = 0.012$$

$$P(\text{review} \mid \text{Non}-\text{spam}) = \frac{2}{6} \quad P(\text{us} \mid \text{Non}-\text{spam}) = \frac{1}{6} \quad P(\text{Non}-\text{spam}) = \frac{2}{6}$$

$$P(\mathbf{x} \mid \text{Non}-\text{spam})\, P(\text{Non}-\text{spam}) = 0.0185$$

Document is likely a non-spam.

Vocabulary	Spam Count	Non-spam Count
send	3	1
us	3	1
your	3	1
password	2	1
review	1	2
account	1	0
	13	6

NB Classifier – Spam Filtering

Naïve Bayes (NB) Classification – Example:

- For $\mathbf{x}$ = "review account".
- For 'account': non-spam count is zero. Consequently, $P(\text{account} \mid \text{Non} - \text{spam}) = 0$.

Solution: Add 1 smoothing

$$P(\text{Spam}) = \frac{4}{6} \qquad P(\text{Non} - \text{spam}) = \frac{2}{6}$$

$$P(\text{review} \mid \text{Spam}) = \frac{1+1}{13+6} = \frac{2}{19} \qquad P(\text{account} \mid \text{Spam}) = \frac{1+1}{13+6} = \frac{2}{19}$$

We have added numerator factor times the size of the vocabulary in the denominator.

$$P(\text{review} \mid \text{Non} - \text{spam}) = \frac{2+1}{6+6} = \frac{3}{12} \qquad P(\text{account} \mid \text{Non} - \text{spam}) = \frac{0+1}{6+6} = \frac{1}{12}$$

$$P(\mathbf{x} \mid \text{Spam})\, P(\text{Spam}) = 0.00738$$

$$P(\mathbf{x} \mid \text{Non} - \text{spam})\, P(\text{Non} - \text{spam}) = 0.00694$$

Document is likely a spam.

Vocabulary	Spam Count	Non-spam Count
send	3	1
us	3	1
your	3	1
password	2	1
review	1	2
account	1	0
	13	6