CS 107, Probability, Spring 2019
Lecture 17

Michael Poghosyan

AUA

25 February 2019

Content

- Naive Bayes Classification

LZ

We have asked two persons to make a coin-tossing experiment (120 tosses) to obtain a random sequence of H, T-s of length 120. The recorded responses are:

LZ

We have asked two persons to make a coin-tossing experiment (120 tosses) to obtain a random sequence of H, T-s of length 120. The recorded responses are:

LZ

$$HHTTHTTHTHHTTHTTTHHTHTHTTHHTHT$$
$$THTHTHHTHHTTHHTHHHTHTTHTTHHTHT$$
$$THHTHTTHHTHTHHTTHHHTTHTHTHTTTH$$
$$THTHTHTHTHHTTHTTHTHHTTHTTHHTHH$$

and

LZ

HHTTHTTHTHHTTHTTTHHTHTHTTHHTHT

THTHTHHTHHTTHHTHHHTHTTHTTHHTHT

THHTHTTHHTHTHHTTHHHTTHTHTHTTTH

THTHTHTHTHHTTHTTHTHHTTHTTHHTHH

and

HHHTTTHHHTHTHHHHHHTHTTTHHTTHHT

HTHTTTHHTHTTHHHTHTTTTHTHHHTHHT

THHHTHHTTHHHHHHTHHTHTHHHTTTTTT

TTHHTHHHTHHHHTHHHHTHTHHTHHTHTH

LZ

HHTTHTTHTHHTTHTTTHHTHTHTTHHTHT

THTHTHHTHHTTHHTHHHTHTTHTTHHTHT

THHTHTTHHTHTHHTTHHHTTHTHTHTTTH

THTHTHTHTHHTTHTTHTHHTTHTTHHTHH

and

HHHTTTHHHTHTHHHHHHTHTTTHHTTHHT

HTHTTTHHTHTTHHHTHTTTTHTHHHTHHT

THHHTHHTTHHHHHHTHHTHTHHHTTTTTT

TTHHTHHHTHHHHTHHHHTHTHHTHHTHTH

One of the persons sent a fake sequence (was too lazy to perform the experiment). Who? Explain!

Naive Bayes Classification

Classification is one of the main topics in Machine Learning, an example of the so-called Supervised Learning Problems.

Naive Bayes Classification

Classification is one of the main topics in Machine Learning, an example of the so-called Supervised Learning Problems. Classification Problem can be stated as follows:

Naive Bayes Classification

Classification is one of the main topics in Machine Learning, an example of the so-called Supervised Learning Problems. Classification Problem can be stated as follows:

- We have a dataset of Observations;

Naive Bayes Classification

Classification is one of the main topics in Machine Learning, an example of the so-called Supervised Learning Problems. Classification Problem can be stated as follows:

- We have a dataset of Observations;
- Each Observation is described, is given through Features;

Naive Bayes Classification

Classification is one of the main topics in Machine Learning, an example of the so-called Supervised Learning Problems. Classification Problem can be stated as follows:

- We have a dataset of Observations;
- Each Observation is described, is given through Features;
- For each Observation from the dataset we know the Label of that Observation;

Naive Bayes Classification

Classification is one of the main topics in Machine Learning, an example of the so-called Supervised Learning Problems. Classification Problem can be stated as follows:

- We have a dataset of Observations;
- Each Observation is described, is given through Features;
- For each Observation from the dataset we know the Label of that Observation;
- The set of Labels is finite

Naive Bayes Classification

Classification is one of the main topics in Machine Learning, an example of the so-called Supervised Learning Problems. Classification Problem can be stated as follows:

- We have a dataset of Observations;
- Each Observation is described, is given through Features;
- For each Observation from the dataset we know the Label of that Observation;
- The set of Labels is finite

The Classification Problem can be stated as: Assume now we have a new Observation described through its Features. Can you predict the Label of that Observation?

Table Form

In the Table Form we can write our problem as:

Table Form

In the Table Form we can write our problem as:

Obs	$Feat_1$	$Feat_2$	...	$Feat_m$	Label
obs_1	$obs_1 f_1$	$obs_1 f_2$	...	$obs_1 f_m$	$obs_1 l$
obs_2	$obs_2 f_1$	$obs_2 f_2$	...	$obs_2 f_m$	$obs_2 l$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
obs_n	$obs_n f_1$	$obs_n f_2$	...	$obs_n f_m$	$obs_n l$

Table Form

In the Table Form we can write our problem as:

Obs	$Feat_1$	$Feat_2$	...	$Feat_m$	Label
obs_1	$obs_1 f_1$	$obs_1 f_2$	...	$obs_1 f_m$	$obs_1 l$
obs_2	$obs_2 f_1$	$obs_2 f_2$	...	$obs_2 f_m$	$obs_2 l$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
obs_n	$obs_n f_1$	$obs_n f_2$	...	$obs_n f_m$	$obs_n l$

Now, we have a new Observation by its Features, and we want to predict the correct Label:

Obs	$Feat_1$	$Feat_2$	...	$Feat_m$	Label
obs	$obsf_1$	$obsf_2$	...	$obsf_m$	?

Examples:

Creditor Rating Example: Here the problem is the following. We have a (historical) list of **Good** and **Bad** (these are our Labels) creditors.

Examples:

Creditor Rating Example: Here the problem is the following. We have a (historical) list of **Good** and **Bad** (these are our Labels) creditors. Each creditor (Observation) is described through the following Features:

Examples:

Creditor Rating Example: Here the problem is the following. We have a (historical) list of **Good** and **Bad** (these are our Labels) creditors. Each creditor (Observation) is described through the following Features:

- Age (in years, $Age \in [20, 80]$),

Examples:

Creditor Rating Example: Here the problem is the following. We have a (historical) list of **Good** and **Bad** (these are our Labels) creditors. Each creditor (Observation) is described through the following Features:

- Age (in years, $Age \in [20, 80]$),
- Wage (in K AMD, $Wage \in [60, 6000]$),

Examples:

Creditor Rating Example: Here the problem is the following. We have a (historical) list of **Good** and **Bad** (these are our Labels) creditors. Each creditor (Observation) is described through the following Features:

- Age (in years, $Age \in [20, 80]$),
- Wage (in K AMD, $Wage \in [60, 6000]$),
- Last Job Duration (in years, shows how long is the person working at his last workplace)

Examples:

Creditor Rating Example: Here the problem is the following. We have a (historical) list of **Good** and **Bad** (these are our Labels) creditors. Each creditor (Observation) is described through the following Features:

- Age (in years, $Age \in [20, 80]$),
- Wage (in K AMD, $Wage \in [60, 6000]$),
- Last Job Duration (in years, shows how long is the person working at his last workplace)
- Sex (f/m)

Examples:

Creditor Rating Example: Here the problem is the following. We have a (historical) list of **Good** and **Bad** (these are our Labels) creditors. Each creditor (Observation) is described through the following Features:

- Age (in years, $Age \in [20, 80]$),
- Wage (in K AMD, $Wage \in [60, 6000]$),
- Last Job Duration (in years, shows how long is the person working at his last workplace)
- Sex (f/m)
- Credit History (y/n, indicates if the Creditor has a Credit History),

Examples:

Creditor Rating Example: Here the problem is the following. We have a (historical) list of **Good** and **Bad** (these are our Labels) creditors. Each creditor (Observation) is described through the following Features:

- Age (in years, $Age \in [20, 80]$),
- Wage (in K AMD, $Wage \in [60, 6000]$),
- Last Job Duration (in years, shows how long is the person working at his last workplace)
- Sex (f/m)
- Credit History (y/n, indicates if the Creditor has a Credit History),
- Number of Late Loan (Re)payments

Examples:

Creditor Rating Example: Here the problem is the following. We have a (historical) list of **Good** and **Bad** (these are our Labels) creditors. Each creditor (Observation) is described through the following Features:

- Age (in years, $Age \in [20, 80]$),
- Wage (in K AMD, $Wage \in [60, 6000]$),
- Last Job Duration (in years, shows how long is the person working at his last workplace)
- Sex (f/m)
- Credit History (y/n, indicates if the Creditor has a Credit History),
- Number of Late Loan (Re)payments
- Credit Amount (in K AMD, in $[100, 5000]$),

Examples:

Say, we can have the following table (of observations):

Examples:

Say, we can have the following table (of observations):

Name	Age	Wage	LJD	Sex	CH	LL	CA	Label
AA	20	80	1.2	M	N	0	1000	G
BB	32	320	5	F	Y	1	500	G
CC	30	140	1	M	Y	0	2300	B
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	

Examples:

Say, we can have the following table (of observations):

Name	Age	Wage	LJD	Sex	CH	LL	CA	Label
AA	20	80	1.2	M	N	0	1000	G
BB	32	320	5	F	Y	1	500	G
CC	30	140	1	M	Y	0	2300	B
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	

Now, assume someone is applying for a new Credit. The Credit Company officer is asking to provide the necessary information, Features. Say, the response is:

Examples:

Say, we can have the following table (of observations):

Name	Age	Wage	LJD	Sex	CH	LL	CA	Label
AA	20	80	1.2	M	N	0	1000	G
BB	32	320	5	F	Y	1	500	G
CC	30	140	1	M	Y	0	2300	B
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Now, assume someone is applying for a new Credit. The Credit Company officer is asking to provide the necessary information, Features. Say, the response is:

Name	Age	Wage	LJD	Sex	CH	LL	CA	Label
KK	25	210	2	F	N	0	3000	?

Our Task is to predict whether the new person will be a Good or Bad Creditor, i.e., will turn the Loan on time or Not?

The Problem Formulation

Now, let us construct the Mathematical Model for our Classification problem: We have

The Problem Formulation

Now, let us construct the Mathematical Model for our Classification problem: We have

- m Features, called $Feat_1, ..., Feat_m$;

The Problem Formulation

Now, let us construct the Mathematical Model for our Classification problem: We have

- m Features, called $Feat_1, ..., Feat_m$;
- Dataset of n Observations, each give by its Features;

The Problem Formulation

Now, let us construct the Mathematical Model for our Classification problem: We have

- m Features, called $Feat_1, ..., Feat_m$;
- Dataset of n Observations, each give by its Features;
- The Set of all Labels;

The Problem Formulation

Now, let us construct the Mathematical Model for our Classification problem: We have

- m Features, called $Feat_1, \dots, Feat_m$;
- Dataset of n Observations, each give by its Features;
- The Set of all Labels;
- We know the correct Labels for our Observations Dataset;

The Problem Formulation

Now, let us construct the Mathematical Model for our Classification problem: We have

- m Features, called $Feat_1, ..., Feat_m$;
- Dataset of n Observations, each give by its Features;
- The Set of all Labels;
- We know the correct Labels for our Observations Dataset;
- Assume Each Feature, and Labels can be anything from some Finite Sets.

The Problem Formulation

Now, let us construct the Mathematical Model for our Classification problem: We have

- m Features, called $Feat_1, ..., Feat_m$;
- Dataset of n Observations, each give by its Features;
- The Set of all Labels;
- We know the correct Labels for our Observations Dataset;
- Assume Each Feature, and Labels can be anything from some Finite Sets. In Statistical terms, we are dealing with Categorical Variables/Features.

The Problem Formulation

Assume:

- $Feat_k = \{f_1^k, f_2^k, ..., f_{p_k}^k\}$, i.e., the k-th Feature can be anything from this finite set.

The Problem Formulation

Assume:

- $Feat_k = \{f_1^k, f_2^k, ..., f_{p_k}^k\}$, i.e., the k-th Feature can be anything from this finite set. Say, $Sex = \{F, M\}$ or $Age = \{20, 21, 22, ..., 80\}$ (we assume that our Features are Discrete and Finite!).
- $Labels = \{\ell_1, \ell_2, ..., \ell_q\}$,

The Problem Formulation

Assume:

- $Feat_k = \{f_1^k, f_2^k, ..., f_{p_k}^k\}$, i.e., the k-th Feature can be anything from this finite set. Say, $Sex = \{F, M\}$ or $Age = \{20, 21, 22, ..., 80\}$ (we assume that our Features are Discrete and Finite!).
- $Labels = \{\ell_1, \ell_2, ..., \ell_q\}$, say, in our Example $Labels = \{Good, Bad\}$.

The Problem Formulation

Assume:

- $Feat_k = \{f_1^k, f_2^k, ..., f_{p_k}^k\}$, i.e., the k-th Feature can be anything from this finite set. Say, $Sex = \{F, M\}$ or $Age = \{20, 21, 22, ..., 80\}$ (we assume that our Features are Discrete and Finite!).
- $Labels = \{\ell_1, \ell_2, ..., \ell_q\}$, say, in our Example $Labels = \{Good, Bad\}$. In other Example, we can have $Labels = \{Dog, Cat, Donkey\}$ etc.
- o_k is our k-th Observation, given by its Features:

$$o_k = (o_k f_1, o_k f_2, ..., o_k f_m), \qquad o_k f_i \in F_i$$

The Problem Formulation

Assume:

- $Feat_k = \{f_1^k, f_2^k, ..., f_{p_k}^k\}$, i.e., the k-th Feature can be anything from this finite set. Say, $Sex = \{F, M\}$ or $Age = \{20, 21, 22, ..., 80\}$ (we assume that our Features are Discrete and Finite!).
- $Labels = \{\ell_1, \ell_2, ..., \ell_q\}$, say, in our Example $Labels = \{Good, Bad\}$. In other Example, we can have $Labels = \{Dog, Cat, Donkey\}$ etc.
- o_k is our k-th Observation, given by its Features:

$$o_k = (o_k f_1, o_k f_2, ..., o_k f_m), \qquad o_k f_i \in F_i$$

and Label $o_k l$

- And we have a new observation $o = (f_1, f_2, ..., f_m)$. We want to predict its Label ℓ.

Mathematical Model

The Correct Label of our new Creditor depends, of course, on chance, is not known in advance.

Mathematical Model

The Correct Label of our new Creditor depends, of course, on chance, is not known in advance. Maybe that person will return the Load on time or not.

Mathematical Model

The Correct Label of our new Creditor depends, of course, on chance, is not known in advance. Maybe that person will return the Load on time or not. And we know the tool to model the uncertainty - Probability (theory, of course)!!
To make a Probabilistic Model, we denote by F_k the k-th Feature of a Random Creditor.

Mathematical Model

The Correct Label of our new Creditor depends, of course, on chance, is not known in advance. Maybe that person will return the Load on time or not. And we know the tool to model the uncertainty - Probability (theory, of course)!!

To make a Probabilistic Model, we denote by F_k the k-th Feature of a Random Creditor. So F_k is random and it is from $Feat_k$ (we assume all values are equiprobable!).

Mathematical Model

The Correct Label of our new Creditor depends, of course, on chance, is not known in advance. Maybe that person will return the Load on time or not. And we know the tool to model the uncertainty - Probability (theory, of course)!!
To make a Probabilistic Model, we denote by F_k the k-th Feature of a Random Creditor. So F_k is random and it is from $Feat_k$ (we assume all values are equiprobable!). And let L be his/her Label, which is random again, from the set *Labels*.

Mathematical Model

The idea of the Naive Bayes Classification is simple:

Mathematical Model

The idea of the Naive Bayes Classification is simple:

Choose (Predict) the Label with the highest probability to appear under given information.

Mathematical Model

The idea of the Naive Bayes Classification is simple:

Choose (Predict) the Label with the highest probability to appear under given information.

Using the above notations, we want to calculate

$$\mathbb{P}(L = \ell_j | F_1 = f_1, F_2 = f_2, ..., F_m = f_m) \quad \text{for} \quad j = 1, ..., q$$

Mathematical Model

The idea of the Naive Bayes Classification is simple:

Choose (Predict) the Label with the highest probability to appear under given information.

Using the above notations, we want to calculate

$$\mathbb{P}(L = \ell_j | F_1 = f_1, F_2 = f_2, ..., F_m = f_m) \quad \text{for} \quad j = 1, ..., q$$

and then choose Label giving the Maximal of these Conditional Probabilities,

Mathematical Model

The idea of the Naive Bayes Classification is simple:

Choose (Predict) the Label with the highest probability to appear under given information.

Using the above notations, we want to calculate

$$\mathbb{P}(L = \ell_j | F_1 = f_1, F_2 = f_2, ..., F_m = f_m) \quad \text{for} \quad j = 1, ..., q$$

and then choose Label giving the Maximal of these Conditional Probabilities, i.e., to find

$$\ell = \underset{j}{argmax}\ \mathbb{P}(L = \ell_j | F_1 = f_1, F_2 = f_2, ..., F_m = f_m)$$

Mathematical Model

So our aim is to calculate

$$\underset{j}{argmax}\ \mathbb{P}(L = \ell_j | F_1 = f_1, F_2 = f_2, ..., F_m = f_m)$$

Mathematical Model

So our aim is to calculate

$$\underset{j}{argmax}\ \mathbb{P}(L = \ell_j | F_1 = f_1, F_2 = f_2, ..., F_m = f_m)$$

Now, we need to calculate these Conditional Probabilities.

Mathematical Model

So our aim is to calculate

$$\underset{j}{argmax}\ \mathbb{P}(L = \ell_j | F_1 = f_1, F_2 = f_2, ..., F_m = f_m)$$

Now, we need to calculate these Conditional Probabilities. And we will use our Good Old Friend Bayes Formula!

Mathematical Model

So our aim is to calculate

$$\underset{j}{argmax}\ \mathbb{P}(L = \ell_j | F_1 = f_1, F_2 = f_2, ..., F_m = f_m)$$

Now, we need to calculate these Conditional Probabilities. And we will use our Good Old Friend Bayes Formula!
By that Bayes Formula,

$$\mathbb{P}(L = \ell_j | F_1 = f_1, F_2 = f_2, ..., F_m = f_m) =$$

$$= \frac{\mathbb{P}(F_1 = f_1, F_2 = f_2, ..., F_m = f_m | L = \ell_j) \cdot \mathbb{P}(L = \ell_j)}{\mathbb{P}(F_1 = f_1, F_2 = f_2, ..., F_m = f_m)}$$

Mathematical Model

So we need to calculate

$$\underset{j}{argmax} \frac{\mathbb{P}(F_1 = f_1, F_2 = f_2, ..., F_m = f_m | L = \ell_j) \cdot \mathbb{P}(L = \ell_j)}{\mathbb{P}(F_1 = f_1, F_2 = f_2, ..., F_m = f_m)}$$

Mathematical Model

So we need to calculate

$$\underset{j}{argmax} \frac{\mathbb{P}(F_1 = f_1, F_2 = f_2, ..., F_m = f_m | L = \ell_j) \cdot \mathbb{P}(L = \ell_j)}{\mathbb{P}(F_1 = f_1, F_2 = f_2, ..., F_m = f_m)}$$

But here the denominator is independent of j!!! Uraa!! We can solve instead:

$$\underset{j}{argmax}\ \mathbb{P}(F_1 = f_1, F_2 = f_2, ..., F_m = f_m | L = \ell_j) \cdot \mathbb{P}(L = \ell_j)$$

Mathematical Model

So we need to calculate

$$\underset{j}{argmax} \; \frac{\mathbb{P}(F_1 = f_1, F_2 = f_2, ..., F_m = f_m | L = \ell_j) \cdot \mathbb{P}(L = \ell_j)}{\mathbb{P}(F_1 = f_1, F_2 = f_2, ..., F_m = f_m)}$$

But here the denominator is independent of j!!! Uraa!! We can solve instead:

$$\underset{j}{argmax} \; \mathbb{P}(F_1 = f_1, F_2 = f_2, ..., F_m = f_m | L = \ell_j) \cdot \mathbb{P}(L = \ell_j)$$

Now, another simplification:

Naive Bayes Classification Method assumes Conditional Independence of Features, i.e. we assume that for any j,

$$\mathbb{P}(F_1 = f_1, F_2 = f_2, ..., F_m = f_m | L = \ell_j) =$$

$$= \mathbb{P}(F_1 = f_1 | L = \ell_j) \cdot \mathbb{P}(F_2 = f_2 | L = \ell_j) \cdot ... \cdot \mathbb{P}(F_m = f_m | L = \ell_j)$$

Mathematical Model

Finally, we have reduced our problem to: find

$$\underset{j}{argmax}\ \mathbb{P}(F_1 = f_1 | L = \ell_j) \cdot \ldots \cdot \mathbb{P}(F_m = f_m | L = \ell_j) \cdot \mathbb{P}(L = \ell_j)$$

Mathematical Model

Finally, we have reduced our problem to: find

$$\underset{j}{argmax}\ \mathbb{P}(F_1 = f_1 | L = \ell_j) \cdot \ldots \cdot \mathbb{P}(F_m = f_m | L = \ell_j) \cdot \mathbb{P}(L = \ell_j)$$

Now, to calculate these Probabilities, we use our dataset:

$$\mathbb{P}(L = \ell_j) = \frac{\#\text{observations with the label } \ell_j}{\#\text{all observations}};$$

Mathematical Model

Finally, we have reduced our problem to: find

$$\underset{j}{argmax}\ \mathbb{P}(F_1 = f_1 | L = \ell_j) \cdot \ldots \cdot \mathbb{P}(F_m = f_m | L = \ell_j) \cdot \mathbb{P}(L = \ell_j)$$

Now, to calculate these Probabilities, we use our dataset:

$$\mathbb{P}(L = \ell_j) = \frac{\#\text{observations with the label } \ell_j}{\#\text{all observations}};$$

$$\mathbb{P}(F_k = f_k | L = \ell_j) = \frac{\#\text{observations with } F_k = f_k \text{ and label } \ell_j}{\#\text{all observations with labels } \ell_j}.$$

The Algorithm:

So the Algorithm is the following:

The Algorithm:

So the Algorithm is the following:

- For any j running over the indices of Labels:

The Algorithm:

So the Algorithm is the following:

- For any j running over the indices of Labels:
    - Calculate $p(L = \ell_j)$,

The Algorithm:

So the Algorithm is the following:

- For any j running over the indices of Labels:
  - Calculate $p(L = \ell_j)$,
  - For any k, calculate $p(F_k = f_k | L = \ell_j)$;

The Algorithm:

So the Algorithm is the following:

- For any j running over the indices of Labels:
  - Calculate $p(L = \ell_j)$,
  - For any k, calculate $p(F_k = f_k | L = \ell_j)$;
  - Calculate the product
    $\mathbb{P}(F_1 = f_1 | L = \ell_j) \cdot ... \cdot \mathbb{P}(F_m = f_m | L = \ell_j) \cdot \mathbb{P}(L = \ell_j)$

The Algorithm:

So the Algorithm is the following:

- For any j running over the indices of Labels:
  - Calculate $p(L = \ell_j)$,
  - For any k, calculate $p(F_k = f_k | L = \ell_j)$;
  - Calculate the product
    $\mathbb{P}(F_1 = f_1 | L = \ell_j) \cdot \ldots \cdot \mathbb{P}(F_m = f_m | L = \ell_j) \cdot \mathbb{P}(L = \ell_j)$
- Find for which Label the obtained product is the maximal

The Algorithm:

So the Algorithm is the following:

- For any j running over the indices of Labels:
  - Calculate $p(L = \ell_j)$,
  - For any k, calculate $p(F_k = f_k | L = \ell_j)$;
  - Calculate the product $\mathbb{P}(F_1 = f_1 | L = \ell_j) \cdot \ldots \cdot \mathbb{P}(F_m = f_m | L = \ell_j) \cdot \mathbb{P}(L = \ell_j)$
- Find for which Label the obtained product is the maximal
- Predict that Label

The Algorithm:

So the Algorithm is the following:

- For any j running over the indices of Labels:
  - Calculate $p(L = \ell_j)$,
  - For any k, calculate $p(F_k = f_k | L = \ell_j)$;
  - Calculate the product $\mathbb{P}(F_1 = f_1 | L = \ell_j) \cdot ... \cdot \mathbb{P}(F_m = f_m | L = \ell_j) \cdot \mathbb{P}(L = \ell_j)$
- Find for which Label the obtained product is the maximal
- Predict that Label
- Wait for the Google or FB Machine Learning Team offer in few days $\ddot\smile$