CS 383/613 – Machine Learning

Probabilistic/Statistical Classification

Slides adapted from material created by E. Alpaydin
Prof. Mordohai, Prof. Greenstadt, Pattern Classification (2nd Ed.),
Pattern Recognition and Machine Learning

Objectives

- Inference
- Bayesian Learning
- Naïve Bayes
- Gaussian Naïve Bayes

Statistical Learning

Statistical Learning

- For our next approach to classification, we will look at the probability and statistics of our labeled data to make predictions on new data
- Review the Prob/Stats Week 0 slides!
- Hopefully, these methods matches some of your intuition about data

Inference

- Our statistical learning will start with the concept of *inference*
  - Given distribution of seen data, what can we *infer* about new data?
- Given evidence/features $\boldsymbol{x} = [x_1, x_2, \ldots, x_D]$ what is the *probability* that our object came from class i?

$$P(y = i | feature_1 = x_1, feature_2 = x_2, \ldots, feature_D = x_D)$$

- We call this the *posterior.*
- And recall from probability that this is a *conditional probability:*

"Given the first feature has value x_1, the second has x_2, etc.. what is the probability that our class was i?"

Inference

- From the rules of probability, we can compute our posterior as:

$$P(y = i | f_1 = x_1, .., f_D = x_D) = \frac{P(y = i, f_1 = x_1, .., f_D = x_D)}{P(f_1 = x_1, \dots f_D = x_D)}$$

- Recall that $P(a, b, c)$ called the *joint probability* and tells us the probability of these things happening at the same time.
- Now we can compute the posterior purely in terms of the joints
- And given enough data we should be able get the joints easily directly from our data!

The Joint Distribution

- How to make a joint distribution:

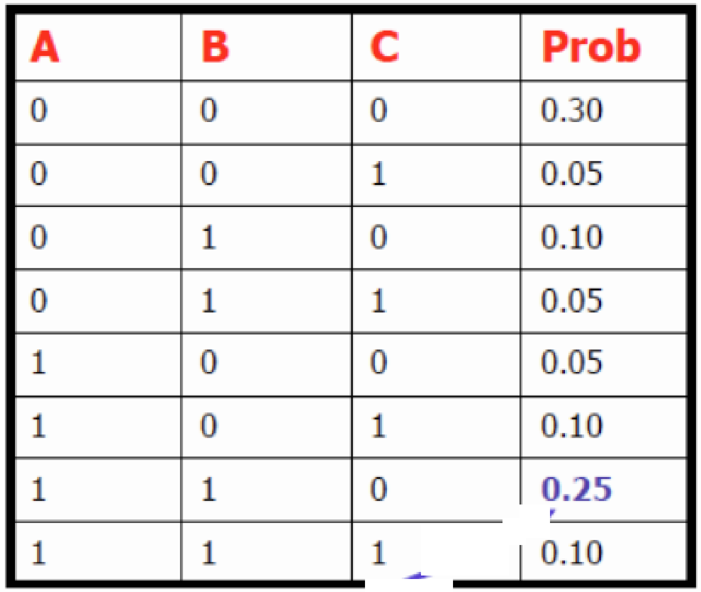
1. Make a table listing all combinations of values of your variables (if there are M Boolean variables, then the table will have 2^M rows)
2. Count how many times in your data each combination occurs
3. Normalize those counts by the total data size in order to arrive a probabilities.

Note: The sum of joints must be equal to one

Learning a Joint Distribution

- To Build a JD (joint distribution) table for your attributes in which the probabilities are unspecified just fill in each row with

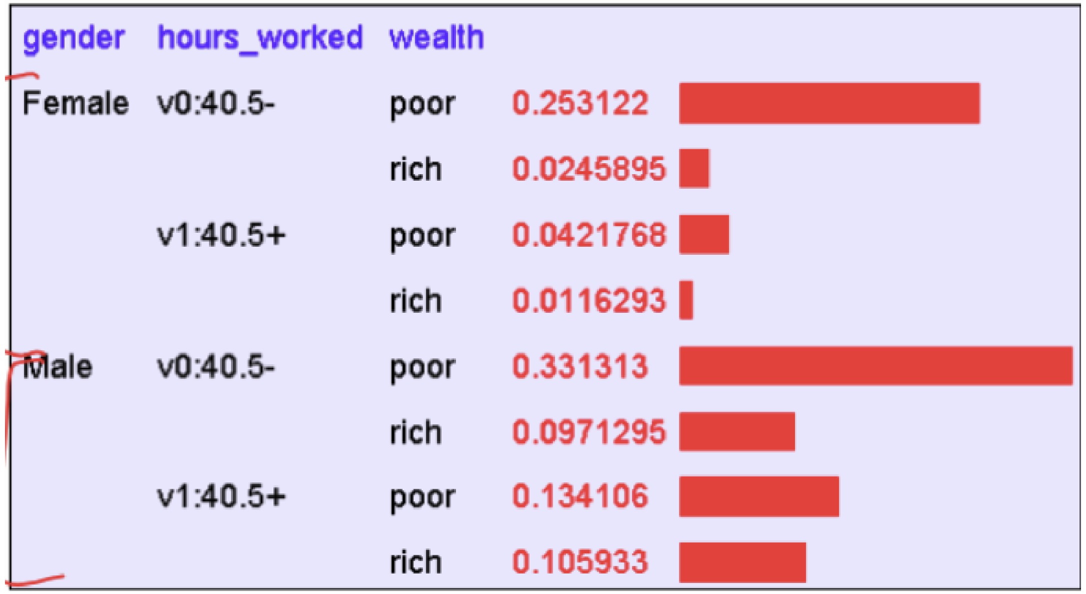
$$P(row) = \frac{records\ matching\ row}{total\ number\ of\ records}$$



A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

Example

- This JD was obtained by learning from three attributes in the UCI "Adult" Census database



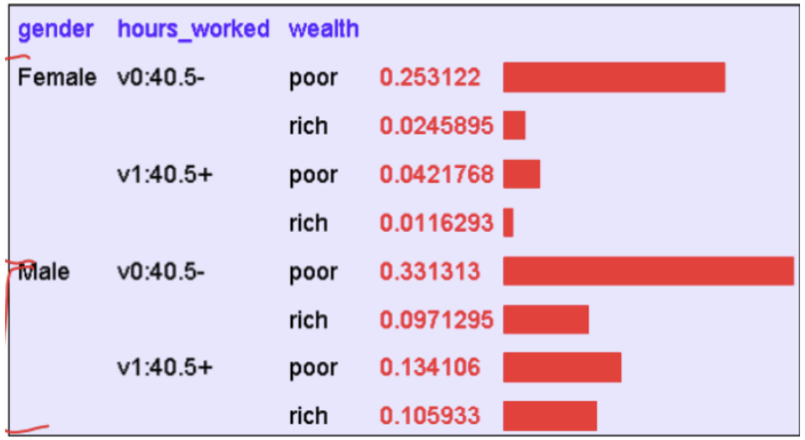
gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Using the Joint

- Once you have the JD you can easily compute the probability of any logical expression involving your attributes.
- One useful law of probability to help do this is the *law of total probability:*

$$P(Y) = \sum_{i} P(Y \cap x_i) = \sum_{rows\,with\,Y} P(row)$$

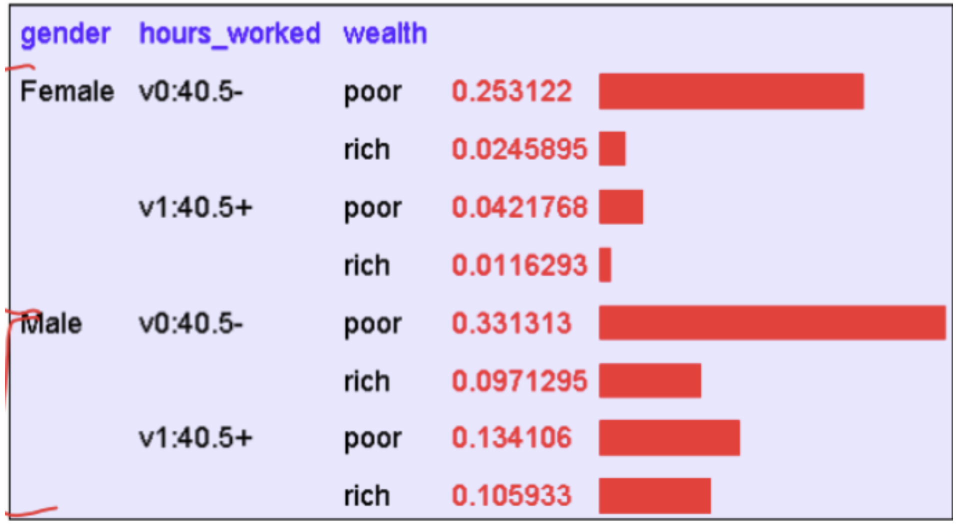
- Examples:
  - What is P(wealth=Poor)?
  - What is P(wealth=Poor, gender=Male)?
- We can also easily compute joint/conditional probabilities using the joint distributions.
- What is $P(gender = Male|wealth = Poor)$?



gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	
		rich	0.0245895	
	v1:40.5+	poor	0.0421768	
		rich	0.0116293	
Male	v0:40.5-	poor	0.331313	
		rich	0.0971295	
	v1:40.5+	poor	0.134106	
		rich	0.105933	

Example

- Suppose we want to figure out given gender and hours worked, what is the wealth?
- What is $P(W = rich | G = female, H = 40.5-)$?



gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Inference is a big deal!

- There’s tons of times you use it:
  - I’ve got this evidence. What’s the chance that my conclusion is true?
  - I’ve got a sore neck. How likely am I to have Meningitis?
  - The lights are out and it’s 9pm. What is the likelihood that my spouse is asleep?

Using Inference for Classification

- How can we use this for classification?
- Consider $\boldsymbol{x}$, a set of D features
- We'll write the posterior shorthand as $P(y = i|\boldsymbol{x})$ and the joint as $P(y = i, \boldsymbol{x})$
- To figure out which class a set of features should belong to we can just choose the class that maximizes the posterior probability

$$\hat{y} = arg\max_{i} P(y = i|\boldsymbol{x})$$

$$= arg\max_{i} \left(\frac{P(y = i, \boldsymbol{x})}{P(x)} \right)$$

Using Inference for Classification

$$\hat{y} = \underset{i}{argmax}\left(\frac{P(y=i,\boldsymbol{x})}{P(\boldsymbol{x})}\right)$$

- But since $P(x)$is the same for all classes we can just do:

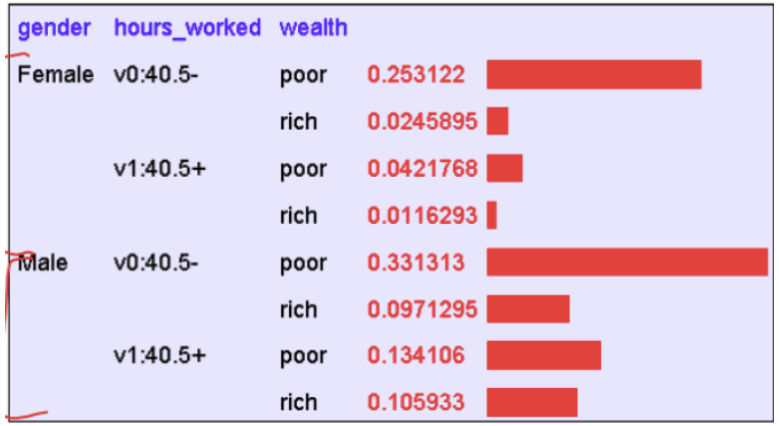
$$\hat{y} = \underset{i}{argmax}\, P(y=i,\boldsymbol{x})$$

- And if we have $P(y=i,\boldsymbol{x})$ for all classes i then you can compute the actual probabilities, $P(y=i|\boldsymbol{x})$ by dividing by the sum of the joint probabilities:
  - Let: $\rho = \sum_{k=1}^{K} P(y=k,\boldsymbol{x})$
  - Now we can compute $P(y=i|x)$ as

$$P(y=i|\boldsymbol{x}) = \frac{P(y=i,\boldsymbol{x})}{\rho}$$

Inference for Classification Example

- Given a rich male let’s classify them as having worked more or less than 40.5 hours per week
  - $P(hours = 40.5 + | gender = male, wealth\ rich)$
    $\propto P(hours = 40.5+, gender = male, wealth = rich)$
  - $P(hours = 40.5 - | gender = male, wealth = rich)$
    $\propto P(hours = 40.5-, gender = male, wealth = rich)$



gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Naïve Bayesian Inference/Classification

Naïve Bayes

- If our data is relatively sparse, then creating a reliable joint distribution $P(\boldsymbol{x}, y)$.
- To overcome this limitation, we can use a combination of Bayes rule, with a naïve assumption that each feature is conditionally independent of one other.
- Together, this approach is known as *Naïve Bayes.*
- Bayes Rule allows us to compute our posterior $P(y|\boldsymbol{x})$ as:

$$P(y|\boldsymbol{x}) = \frac{P(y)P(\boldsymbol{x}|y)}{P(\boldsymbol{x})}$$

  - Where $P(y)$ is known as the class prior (the probability of the class occurring, independent of any observation).
  - And where $P(\boldsymbol{x}|y)$ is known as the *generative likelihood.*

Naïve Bayes

- The naïve assumption that each feature is conditionally independent on one another (that is, that given one feature conditioned on our class we can't say anything about the other features), allows us to write the generative likelihood as:

$$P(\boldsymbol{x}|y) \approx \prod_{j=1}^{D} P\left(x_j|y\right)$$

- This is a large assumption, but one that is often made.
- So, we can now approximate our posterior as:

$$P(y|\boldsymbol{x}) = \frac{P(y)P(\boldsymbol{x}|y)}{P(\boldsymbol{x})} \approx \frac{P(y)\prod_{j=1}^{D} P\left(x_j|y\right)}{P(\boldsymbol{x})}$$

Naïve Bayes Probability

$$P(y|\boldsymbol{x}) = \frac{P(y)P(\boldsymbol{x}|y)}{P(\boldsymbol{x})} \approx \frac{P(y)\prod_{j=1}^{D} P\left(x_j|y\right)}{P(\boldsymbol{x})}$$

- Of course, we likely can't easily compute the evidence $P(\boldsymbol{x})$
- But the good news is that we know the probabilities over the classes should sum to one, so we can just divide by the sum of their numerators (again)!
- If we have K classes, we can let:

$$\rho = \sum_{k=1}^{K}\left(P(y=k)\prod_{j=1}^{D} P\left(x_j|y=k\right)\right)$$

- And now:

$$P(y=i|\boldsymbol{x}) = \frac{P(y=i)\prod_{j=1}^{D} P\left(x_j|y=i\right)}{\rho}$$

Log-Exponent Trick

$$P(y|\boldsymbol{x}) \approx \frac{P(y)\prod_{j=1}^{D} P\left(x_j\middle|y\right)}{P(\boldsymbol{x})}$$

- Since $0 \leq P\left(x_j\middle|y\right) \leq 1$, if we have a lot of features, there’s the potential for underflow.
- Therefore, it is common to use the *log-exponent trick* to avoid underflows.

$$\log(P(y|\boldsymbol{x}) \approx \log\left(P(y)\right) + \sum_{j=1}^{D} \log(P\left(x_j\middle|y\right) - \log\left(P(\boldsymbol{x})\right)$$

- If all we care about it the most likely class, we can just choose:

$$\hat{y} = argmax_i\left(\log\left(P(y=i)\right) + \sum_{j=1}^{D} \log(P\left(x_j\middle|y=i\right)\right)$$

- If we need true probabilities, we can always “undo” the log after the fact:

$$P(y|\boldsymbol{x}) = e^{\log\left(P(y|\boldsymbol{x})\right)}$$

- One last thing….
  - If you decided to work in logs, be careful of $\log(0)$!

Example

- Let’s try to determine if an object is a banana, an orange, or something else based on its length, sweetness, and color.
  - Where length, sweetness and color are all binary features (isLong?, isSweet?, isYellow?)
- Below are tables describing attributes of three types of fruit (over 1000 samples).

Banana			
Long	*Sweet*	*Yellow*	*Count*
F	F	F	50
F	F	T	50
F	T	F	0
F	T	T	0
T	F	F	0
T	F	T	50
T	T	F	0
T	T	T	350

Orange			
Long	*Sweet*	*Yellow*	*Count*
F	F	F	0
F	F	T	150
F	T	F	0
F	T	T	150
T	F	F	0
T	F	T	0
T	T	F	0
T	T	T	0

Other			
Long	*Sweet*	*Yellow*	*Count*
F	F	F	0
F	F	T	0
F	T	F	50
F	T	T	50
T	F	F	50
T	F	T	0
T	T	F	50
T	T	T	0

Example

- Let’s compute the probability of being a banana if we observe that a fruit is long, not sweet, and yellow.

$$P(B|L,\neg S,Y)$$

- First let’s try to do this using inference:

$$P(B|L,\neg S,Y)=\frac{P(B,L,\neg S,Y)}{P(L,\neg S,Y)}$$

Banana			
Long	*Sweet*	*Yellow*	*Count*
F	F	F	50
F	F	T	50
F	T	F	0
F	T	T	0
T	F	F	0
T	F	T	50
T	T	F	0
T	T	T	350

Orange			
Long	*Sweet*	*Yellow*	*Count*
F	F	F	0
F	F	T	150
F	T	F	0
F	T	T	150
T	F	F	0
T	F	T	0
T	T	F	0
T	T	T	0

Other			
Long	*Sweet*	*Yellow*	*Count*
F	F	F	0
F	F	T	0
F	T	F	50
F	T	T	50
T	F	F	50
T	F	T	0
T	T	F	50
T	T	T	0

Example

- How about with Naïve Bayes?

$$P(B|L,\neg S,Y) \approx\propto P(B)P(L|B)P(\neg S|B)P(Y|B)$$

- But if I want an actual probability, we also need to compute this for the Orange and Other class than divide by their sum…

Banana			
Long	*Sweet*	*Yellow*	*Count*
F	F	F	50
F	F	T	50
F	T	F	0
F	T	T	0
T	F	F	0
T	F	T	50
T	T	F	0
T	T	T	350

Orange			
Long	*Sweet*	*Yellow*	*Count*
F	F	F	0
F	F	T	150
F	T	F	0
F	T	T	150
T	F	F	0
T	F	T	0
T	T	F	0
T	T	T	0

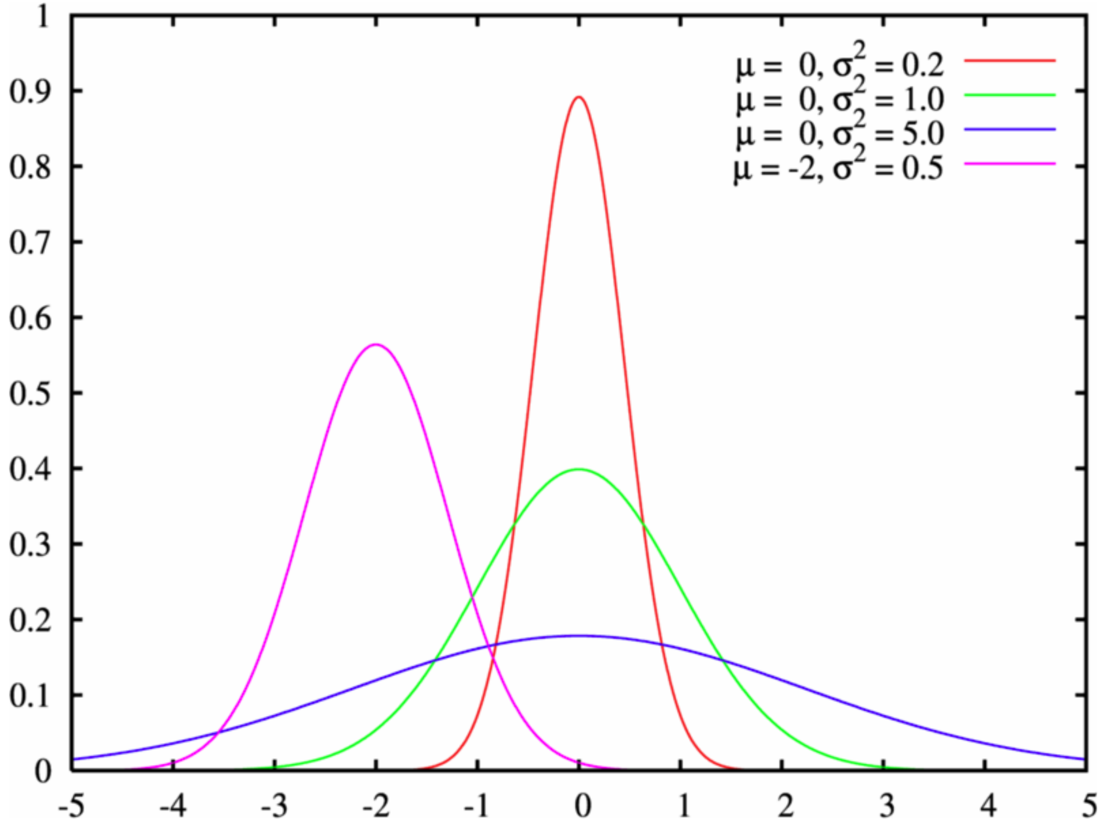
Other			
Long	*Sweet*	*Yellow*	*Count*
F	F	F	0
F	F	T	0
F	T	F	50
F	T	T	50
T	F	F	50
T	F	T	0
T	T	F	50
T	T	T	0

Continuous Valued Data

- This is another example of algorithms that work most easily/naturally with categorical features.
- So again, we could convert continuous ones to categorical ordinal.
- However, it is common to use Naïve Bayes in conjunction with the assumption that continuous features following a Gaussian (normal) distribution, in which case we can perform Gaussian Naïve Bayes using continuous features!
- Let’s dig into this!

Gaussian Distribution

- Recall a Gaussian distribution, also known as
  - Normal distribution
  - “Bell shaped curve”
- These can be parameterized by two parameters:
  - Mean μ
  - Standard deviation σ (or variance, σ^2)



Probability Density Functions

- Known distributions have *probability density functions (pdfs),* that, when evaluated for a value, provide a value that can be interpreted as providing a *relatively likelihood.*
  - *Note: A pdf integrated from negative infinity to positive infinity is one. Thus the "relative" part.*
- Let $P(x|y)$ be our actual likelihood of x given y. A pdf $p(x|y)$ then gives us:

$$P(x|y) \propto p(x|y)$$

- Recall that the likelihood can be useful, when combined with Bayes rule, for computing our posterior $P(y|x)$

Normal (Gaussian) Distribution

- The probability density function for a normal (Gaussian) distribution is:

$$P(x|\mu,\sigma) \propto p(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- How can we use this for Naïve Bayes?

Naïve Bayes

- For each class, compute the mean and standard deviation for each feature as $\mu_j^{(y)}$ and $\sigma_j^{(y)}$, respectively.
- Our Norm PDF is then just:

$$P\left(x_j \middle| \mu_j^{(y)}, \sigma_j^{(y)}\right) \propto \frac{1}{\sigma_j^{(y)}\sqrt{2\pi}} e^{-\frac{\left(x_j - \mu_j^{(y)}\right)^2}{2\left(\sigma_j^{(y)}\right)^2}}$$

- We can then compute a posterior probability for class $y = k$ as:

$$P(y = k|x) = \frac{P(y = k)\prod_{j=1}^{D} p\left(x_j \middle| u_j^{(k)}, \sigma_j^{(k)}\right)}{\sum_{i=1}^{K} P(y = i)\prod_{j=1}^{D} p\left(x_j \middle| u_j^{(i)}, \sigma_j^{(i)}\right)}$$

Additional Resources

- https://stackoverflow.com/questions/10059594/a-simple-explanation-of-naive-bayes-classification/20556654#20556654