COMP3308/3608, Lecture 6a

ARTIFICIAL INTELLIGENCE

Statistical-Based Learning (Naïve Bayes)

Reference: Witten, Frank, Hall and Pal, ch.4.2: p.96-105

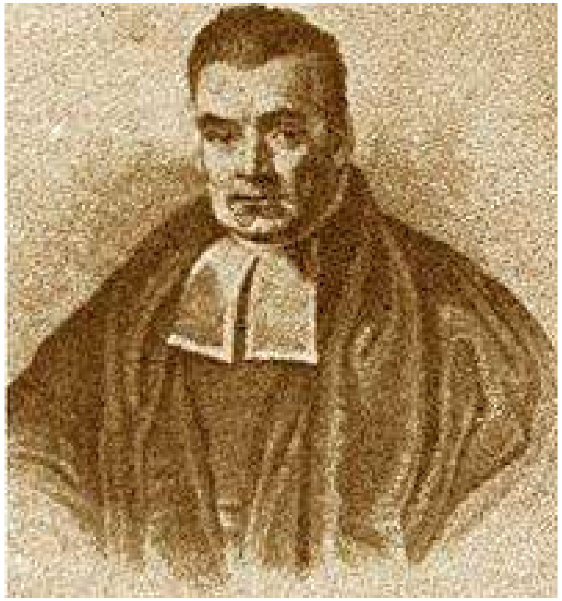
Russell and Norvig, ch.20: p.802-810

Outline

- **Bayes theorem**
- **Naïve Bayes algorithm**
- **Naïve Bayes - issues**
  - **Zero probabilities - Laplace correction**
  - **Dealing with missing values**
  - **Dealing with numeric attributes**

What is Bayesian Classification?

- Bayesian classifiers are statistical classifiers
- They can predict the class membership probability, i.e. the probability that a given example belongs to a particular class
- They are based on the Bayes Theorem



Thomas Bayes (1702-1761)

Bayes Theorem

- **Given a hypothesis *H* and evidence *E* for this hypothesis, then the probability of *H* given *E*, is:**

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

- **Example: Given are instances of fruits, described by their color and shape. Let:**
  - ***E* is red and round**
  - ***H* is the hypothesis that *E* is an apple**

- **What are:**
  - $P(H|E)=?$
  - $P(H)=?$
  - $P(E|H)=?$
  - $P(E)=?$

Bayes Theorem – Example (cont. 1)

- *$P(H|E)$* is the probability that *E* is an apple, given that we have seen that *E* is red and round
  - Called *posteriori probability* of *H* conditioned on *E*

- *$P(H)$* is the probability that any given example is an apple, regardless of how it looks
  - Called *prior probability* of *H*

- The posteriori probability is based on more information that the prior probability which is independent of *E*

Bayes Theorem – Example (cont. 2)

- What is *P(E|H)*?
  - the posteriori probability of *E* conditioned on *H*
  - the probability that *E* is red and round, given that we know that *E* is an apple
- What is *P(E)*?
  - the prior probability of *E*
  - The probability that an example from the fruit data set is red and round

Bayes Theorem for Problem Solving

- **Given: A doctor knows that**
  - **Meningitis causes stiff neck 50% of the time**
  - **Prior probability of any patient having meningitis is 1/50 000**
  - **Prior probability of any patient having stiff neck is 1/20**
- **If a patient has a stiff neck, what is the probability that he has meningitis?**

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

Bayes Theorem for Problem Solving - Answer

- **Given: A doctor knows that**
  - **Meningitis causes stiff neck 50% of the time** $P(S \mid M)$
  - **Prior probability of any patient having meningitis is 1/50 000** $P(M)$
  - **Prior probability of any patient having stiff neck is 1/20** $P(S)$
- **If a patient has a stiff neck, what is the probability that he has meningitis?** $P(M \mid S) = ?$

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5\,(1/50000)}{1/20} = 0.0002$$

Naïve Bayes Algorithm

- The Bayes Theorem can be applied for classification tasks = Naïve Bayes algorithm
- While 1R makes decisions based on a single attribute, Naive Bayes uses all attributes and allows them to make contributions to the decision that are ***equally important*** **and** ***independent*** of one another
- **Assumptions of the Naïve Bayes algorithm**
  - 1) Independence assumption – (the values of the) attributes are conditionally independent of each other, given the class (i.e. for each class value)
  - 2) Equally importance assumption – all attributes are equally important
- Unrealistic assumptions! => it is called ***Naive*** Bayes
  - Attributes are dependent of one another
  - Attributes are not equally important
- But these assumptions lead to a simple method which works surprisingly well in practice!

Naive Bayes on the Weather Example

- **Given: the weather data →**

outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

- **Task: use Naïve Bayes to predict the class (*yes* or no) of the new example**

  `outlook=sunny, temperature=cool, humidity=high, windy=true`

- **The Bayes Theorem:** $P(H \mid E) = \dfrac{P(E \mid H)P(H)}{P(E)}$

- **What are H and E?**
  - **the evidence E is the new example**
  - **the hypothesis H is `play=yes` (and there is another H: `play=no`)**

- **How to use the Bayes Theorem for classification?**
- **Calculate *P(H|E)* for each *H* (class), i.e. P(yes|E) and P(no|E)**
- **Compare them and assign *E* to the class with the highest probability**
- **OK, but for *P(H|E)* we need to calculate *P(E), P(H)* and *P(E|H)* – how to do this? From the given data (this is the training phase of the classifier)**

Naive Bayes on the Weather Example (2)

- We need to calculate and compare P(yes|E) and P(no|E)

$$P(yes \mid E) = \frac{P(E \mid yes)P(yes)}{P(E)}$$

$$P(no \mid E) = \frac{P(E \mid no)P(no)}{P(E)}$$

where E

`outlook=sunny,temperature=cool,`
`humidity=high, windy=true`

1) How to calculate P(E|yes) and P(E|no) ?

Let's split the evidence E into 4 smaller pieces of evidence:

- E1 = `outlook=sunny,` E2 = `temperature=cool`
- E3 = `humidity=high,` E4 = `windy=true`

Let's use the Naïve Bayes's independence assumption: E1, E2, E3 and E4 are *independent* given the class. Then, their combined probability is obtained by multiplication of per-attribute probabilities:

$$P(E \mid yes) = P(E_1 \mid yes)\,P(E_2 \mid yes)\,P(E_3 \mid yes)\,P(E_4 \mid yes)$$

$$P(E \mid no) = P(E_1 \mid no)\,P(E_2 \mid no)\,P(E_3 \mid no)\,P(E_4 \mid no)$$

Naive Bayes on the Weather Example (3)

- **Hence:**

$$P(yes \mid E) = \frac{P(E_1 \mid yes)\,P(E_2 \mid yes)\,P(E_3 \mid yes)\,P(E_4 \mid yes)\,P(yes)}{P(E)}$$

$$P(no \mid E) = \frac{P(E_1 \mid no)\,P(E_2 \mid no)\,P(E_3 \mid no)\,P(E_4 \mid no)\,P(no)}{P(E)}$$

- **In summary:**
  - **Numerator - the probabilities will be estimated from the data**
  - **Denominator – the two denominators are the same (P(E)) and since we are comparing the two fractions, we can just compare the numerators => there is no need to calculate P(E)**

Calculating the Probabilities from the Training Data

E1 = `outlook=sunny`, **E2 =** `temperature=cool`

E3 = `humidity=high`, **E4 =** `windy=true`

$$P(yes \mid E) = \frac{P(E_1 \mid yes)\,P(E_2 \mid yes)\,P(E_3 \mid yes)\,P(E_4 \mid yes)\,P(yes)}{P(E)}$$

- **P(E1|yes)=P(outlook=sunny|yes)=?**
- **P(E2|yes)=P(temp=cool|yes)=?**
- **P(E3|yes)=P(humidity=high|yes)=?**
- **P(E4|yes)=P(windy=true|yes)=?**
- **P(yes)=?**

outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Calculating the Probabilities from the Training Data

E1 = `outlook=sunny`, **E2 =** `temperature=cool`

E3 = `humidity=high`, **E4 =** `windy=true`

$$P(yes \mid E) = \frac{P(E_1 \mid yes)\,P(E_2 \mid yes)\,P(E_3 \mid yes)\,P(E_4 \mid yes)\,P(yes)}{P(E)}$$

- **P(E1|yes)=P(outlook=sunny|yes) = ?/9 = 2/9**
- **P(E2|yes)=P(temp=cool|yes)=?**
- **P(E3|yes)=P(humidity=high|yes)=?**
- **P(E4|yes)=P(windy=true|yes)=?**
- **P(yes)=?**

outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Calculation the Probabilities (2)

- **Weather data - counts and probabilities:**

	outlook		temperature			humidity			windy			play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

proportions of days when play is yes

proportions of days when humidity is normal and play is yes i.e. the probability of humidity to be normal given that play=yes

Calculation the Probabilities (3)

$P(yes \mid E) = ?$

$$P(yes \mid E) = \frac{P(E_1 \mid yes)P(E_2 \mid yes)P(E_3 \mid yes)P(E_4 \mid yes)P(yes)}{P(E)}$$

	outlook		temperature			humidity			windy			play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

$\Rightarrow$ **$P(E_1|yes)=P(outlook=sunny|yes)=2/9$**

$P(E_2|yes)=P(temperature=cool|yes)=3/9$

$P(E_3|yes)=P(humidity=high|yes)=3/9$

$P(E_4|yes)=P(windy=true|yes)=3/9$

- **P(yes) =? - the probability of a yes without knowing any E, i.e. anything about the particular day; the prior probability of yes; P(yes) = 9/14**

Final Calculations

- **By substituting the respective evidence probabilities:**

$$P(yes \mid E) = \frac{\frac{2}{9}\frac{3}{9}\frac{3}{9}\frac{3}{9}\frac{9}{14}}{P(E)} = \frac{0.0053}{P(E)}$$

- **Similarly calculating** $P(no \mid E)$ **:**

$$P(no \mid E) = \frac{\frac{3}{5}\frac{1}{5}\frac{4}{5}\frac{3}{5}\frac{5}{14}}{P(E)} = \frac{0.0206}{P(E)}$$

- => $P(no \mid E) > P(yes \mid E)$
- **=> for the new day `play = no` is more likely than `play = yes`**

Another Example

- **Use the NB classifier to solve the following problem:**

- **Consider a volleyball game between team A and team B**
  - **Team A has won 65% of the time and team B has won 35%**
  - **Among the games won by team A, 30% were when playing on team B's court**
  - **Among the games won by team B, 75% were when playing at home**
- **If team B is hosting the next match, which team is most likely to win?**

Solution

- **host – the team hosting the match {A, B}**
- **winner – the winner of the match {A, B}**

- **Using NB, the task is to compute and compare 2 probabilities:**

  P(winner=A|host=B)

  P(winner=B|host=B)

$$P(winner = A \mid host = B) = \frac{P(host = B \mid winner = A)P(winner = A)}{P(host = B)}$$

$$P(winner = B \mid host = B) = \frac{P(host = B \mid winner = B)P(winner = B)}{P(host = B)}$$

Solution (2)

$$P(winner = A \mid host = B) = \frac{P(host = B \mid winner = A)P(winner = A)}{P(host = B)}$$

$$P(winner = B \mid host = B) = \frac{P(host = B \mid winner = B)P(winner = B)}{P(host = B)}$$

- **Do we know these probabilities:**
  - **P(winner=A)= ? //probability that A wins**
  - **P(winner=B)=? //probability that B wins**
  - **P(host=B|winner=A)=? //probability that team B hosted the match, given that team A won**
  - **P(host=B|winner=B)=? //probability that team B hosted the match, given that team B won**

Solution (3)

$$P(winner = A \mid host = B) = \frac{P(host = B \mid winner = A) P(winner = A)}{P(host = B)}$$

$$P(winner = B \mid host = B) = \frac{P(host = B \mid winner = B) P(winner = B)}{P(host = B)}$$

- **Do we know these probabilities:**
  - **P(winner=A)= ? //probability that A wins =0.65**
  - **P(winner=B)=? //probability that B wins =0.35**
  - **P(host=B|winner=A)=? //probability that team B hosted the match, given that team A won =0.30**
  - **P(host=B|winner=B)=? //probability that team B hosted the match, given that team B won =0.75**

Solution (4)

$$P(winner = A \mid host = B) = \frac{P(host = B \mid winner = A)P(winner = A)}{P(host = B)} =$$

$$= \frac{0.3 * 0.65}{P(host = B)} = 0.195$$

$$P(winner = B \mid host = B) = \frac{P(host = B \mid winner = B)P(winner = B)}{P(host = B)} =$$

$$= \frac{0.75 * 0.35}{P(host = B)} = 0.2625$$

=>NB predicts team B

i.e. NB predicts that if team B is hosting the next match, then team B is more likely to win

Three More Things About Naïve Bayes

- **How to deal with probability values of zero in the numerator?**
- **How do deal with missing values?**
- **How to deal with numeric attributes?**

Problem – Probability Values of 0

	yes	…
sunny	0	…
overcast	4	…
rainy	3	…
		…
sunny	0/9	…
overcast	4/9	…
rainy	3/9	…

- **Suppose that the training data was different:**

`outlook=sunny` had <u>always</u> occurred together with `play=no` (i.e. `outlook=sunny` had <u>never</u> occurred together with `play=yes`)

- **Then:**

  P(outlook=sunny|yes)=0 and

  P(outlook=sunny|no)=1

$$P(yes \mid E) = \frac{\underbrace{P(E_1 \mid yes)}_{=0} P(E_2 \mid yes) P(E_3 \mid yes) P(E_4 \mid yes) P(yes)}{P(E)}$$

  - **=> final probability P(yes|E)=0 no matter of the other probabilities**
- **This is not good!**
  - **The other probabilities are completely ignored due to the multiplication with 0**
  - **I.e. the prediction for new examples with `outlook=sunny` will always be `no`, regardless of the values of the other attributes**

A Simple Trick to Avoid This Problem

- **Assume that our training data is so large that adding 1 to each count would not make difference in calculating the probabilities …**
- **but it will avoid the case of 0 probability**
- **This is called the Laplace correction or Laplace estimator**

"What we know is not much. What we do not know is immense."

- ‘

Pierre-Simon Laplace (1749-1827)

Image from http://en.wikipedia.org/wiki/File:Pierre-Simon_Laplace.jpg

Laplace Correction

- **Add 1 to the numerator and *k* to the denominator, where *k* is the number of attribute values for the given attribute**
- **Example:**
  - **A dataset with 2000 examples, 2 classes: *buy_Mercedes=yes* and *buy_Mercedes=no;* 1000 examples in each class**
  - **1 of the attributes is *income* with 3 values: *low, medium* and *high***
  - **For class *buy_Mercedes=yes* , there are 0 examples with *income=low*, 10 with *income=medium* and 990 with *income=high***
- **Probabilities without the Laplace correction for class *yes:***

  0/1000=0, 10/1000=0.01, 990/1000=0.99
- **Probabilities with the Laplace correction:**

  1/1003=0.001, 11/1003=0.011, 991/1003=0.988
- **The correct probabilities are close to the adjusted probabilities, yet the 0 probability value is avoided!**

Laplace Correction – Modified Weather Example

	yes	...
sunny	0	...
overcast	4	...
rainy	3	...
		...
sunny	0/9	...
overcast	4/9	...
rainy	3/9	...

$$P(sunny \mid yes) = \frac{0+1}{9+3} = \frac{1}{12}$$

$$P(overcast \mid yes) = \frac{4+1}{9+3} = \frac{5}{12}$$

$$P(rainy \mid yes) = \frac{3+1}{9+3} = \frac{4}{12}$$

P(sunny|yes)=0/9 →problem

P(overcast|yes)=4/9

P(rainy|yes)=3/9

Laplace correction

- **Assumes that there are 3 more examples from class *yes*, 1 for each value of *outlook***
- **This results in adding 1 to the numerator and 3 to the denominator of all probabilities**
- **Ensures that an attribute value which occurs 0 times will receive a nonzero (although small) probability**

Generalization of the Laplace Correction: M-estimate

- **Add a small constant m to each denominator and mp_i to each numerator, where p_i is the prior probability of the i values of the attribute:**

$$P(sunny \mid yes) = \frac{2 + mp_1}{9 + m} \qquad P(overcast \mid yes) = \frac{4 + mp_2}{9 + m} \qquad P(rainy \mid yes) = \frac{3 + mp_3}{9 + m}$$

- **Note that $p_1+p_2+\ldots+p_n=1$, n - number of attribute values**
- **Advantage of using prior probabilities – it is rigorous**
- **Disadvantage – computationally expensive to estimate prior probabilities**
- **Large m - the prior probabilities are very important compared with the new evidence coming in from the training data; small m - less important**

- **Typically we assume that each attribute value is equally probable, i.e. $p_1=p_2=\ldots=p_n=1/n$**
- **The Laplace correction is a special case of the m-estimate, where $p_1=p_2=\ldots=p_n=1/n$ and $m=n$. Thus, 1 is added to the numerator and m to the denominator.**

Handling Missing Values - Easy

- **Missing attribute value in the new example – do not include this attribute**
  - **e.g. outlook=?, temperature=cool, humidity=high, windy=true**
  - **Then:** $$P(yes \mid E) = \frac{\frac{3}{9}\frac{3}{9}\frac{3}{9}\frac{9}{14}}{P(E)} = \frac{0.0238}{P(E)} \qquad P(no \mid E) = \frac{\frac{1}{5}\frac{4}{5}\frac{3}{5}\frac{5}{14}}{P(E)} = \frac{0.0343}{P(E)}$$
  - **outlook is not included. Compare these results with the previous results!**
  - **As one of the fractions is missing, the probabilities are higher but the comparison is fair - there is a missing fraction in both cases**

- **Missing attribute value in a training example – do not include this value in the counts**
  - **Calculate the probabilities based on the number of values that actually occur (are not missing) and not on the total number of training examples**

Handling Numeric Attributes

outlook			temperature			humidity			windy			play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3		83	85		86	85	false	6	2	9	5
overcast	4	0		70	80		96	90	true	3	3		
rainy	3	2		68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
				81			75						
sunny	2/9	3/5	*mean*	73	74.6	*mean*	79.1	86.2	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	*std dev*	6.2	7.9	*std dev*	10.2	9.7	true	3/9	3/5		
rainy	3/9	2/5											

numeric

- **We would like to classify the following new example:**

outlook=sunny, temperature=66, humidity=90, windy=true

- **Question: How to calculate**

P(temperature=66|yes)=?, P(humidity=90|yes)=?

P(temperature=66|no)=?, P(humidity=90|no) ?

Using Probability Density Function

- **Answer: By assuming that numerical values have a *normal* (Gaussian, bell curve) probability distribution and using the <u>probability density function</u>**
- **For a *normal* distribution with mean μ and standard deviation σ, the probability density function is:**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

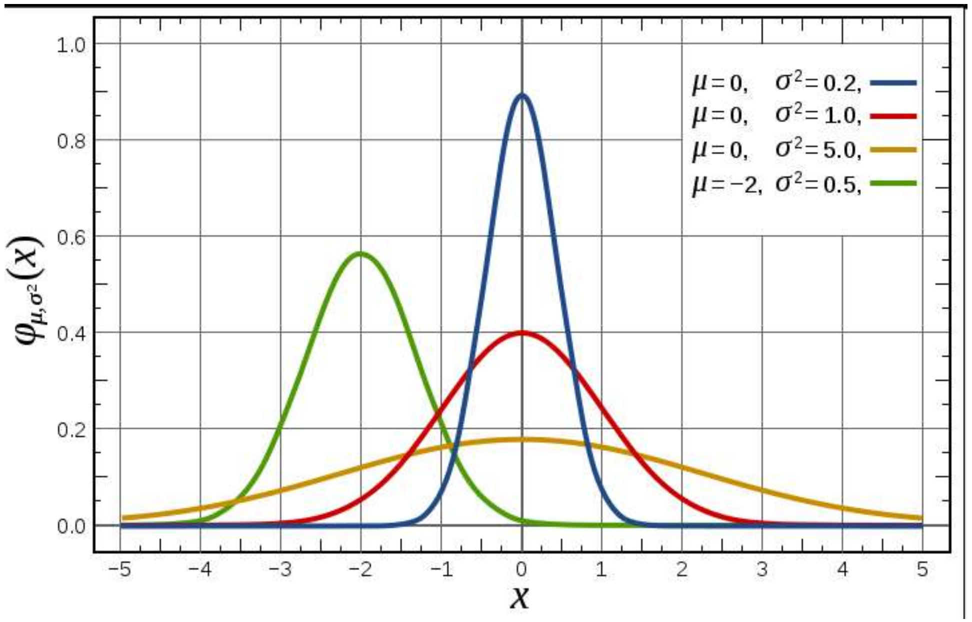


Image from http://en.wikipedia.org/wiki/File:Normal_Distribution_PDF.svg

More on Probability Density Functions

What is the meaning of the probability density function of a continuous random variable?

- **closely related to probability but not exactly the probability (e.g. the probability that *x is <u>exactly</u> 66 is 0)***
- **= the probability that a given value $x \in (x- \varepsilon/2,\ x + \varepsilon/2)$ is $\varepsilon * f(x)$**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Calculating Probabilities Using Probability Density Function

mean for temp. for class=yes

$$f(temperature = 66 \mid yes) = \frac{1}{6.2\sqrt{2\pi}} e^{-\frac{(66-73)^2}{2*6.2^2}} = 0.034$$

std.dev. for temp. for class=yes

$$f(humidity = 90 \mid yes) = 0.0221$$

$$P(yes \mid E) = \frac{\frac{2}{9}\ 0.034\ 0.0221 \frac{3}{9}\frac{9}{14}}{P(E)} = \frac{0.000036}{P(E)}$$

$$P(no \mid E) = \frac{\frac{3}{5}\ 0.0291\ 0.038 \frac{3}{5}\frac{5}{14}}{P(E)} = \frac{0.000136}{P(E)}$$

=>P(no|E) > P(yes|E)

=> no play

- **Compare with the categorical weather data!**

Mean and Standard Deviation - Reminder

- **A reminder how to calculate the mean value μ and standard deviation σ:**

 X is a random variable with values, $x_1, x_2, \ldots x_n$

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n-1}}$$

Note that the denominator is *n-1* not *n*

Naive Bayes - Advantages

- Simple approach – the probabilities are easily computed due to the independence assumption
- Clear semantics for representing, using and learning probabilistic knowledge
- Excellent computational complexity
  - Requires 1 scan of the training data to calculate all statistics (for both nominal and continuous attributes assuming normal distribution):
  - O(pk), p - # training examples, k-valued attributes

- In many cases outperforms more sophisticated learning methods => always try the simple method first!
- Robust to isolated noise points as such points are averaged when estimating the conditional probabilities from data

Naive Bayes - Disadvantages

- **Correlated attributes reduce the power of Naïve Bayes**
  - **Violation of the independence assumption**
  - **Solution: apply feature selection beforehand to identify and discard correlated (redundant) attributes**

- **Normal distribution assumption for numeric attributes - many features are not normally distributed – solutions:**
  - **Discretize the data first, i.e. numerical -> nominal attributes**
  - **Use other probability density functions, e.g. Poisson, binomial, gamma, etc.**
  - **Transform the attribute using a suitable transformation into a normally distributed one (sometimes possible)**
  - **Use kernel density estimation – doesn't assume any particular distribution**