Bayesian Learning

Generative and discriminative classifiers: Naive Bayes
and logistic regression

Conditional Probability

- Probability of an event given the occurrence of some other event.

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{P(X, Y)}{P(Y)}$$

Example

You've been keeping track of the last 1000 emails you received. You find that 100 of them are spam. You also find that 200 of them were put in your junk folder, of which 90 were spam.

Example

You've been keeping track of the last 1000 emails you received. You find that 100 of them are spam. You also find that 200 of them were put in your junk folder, of which 90 were spam.

What is the probability an email you receive is spam?

Example

You've been keeping track of the last 1000 emails you received. You find that 100 of them are spam. You also find that 200 of them were put in your junk folder, of which 90 were spam.

What is the probability an email you receive is spam?

$$P(X) = 100/1000 = .1$$

Example

You've been keeping track of the last 1000 emails you received. You find that 100 of them are spam. You also find that 200 of them were put in your junk folder, of which 90 were spam.

What is the probability an email you receive is spam?

$$P(X) = 100/1000 = .1$$

What is the probability an email you receive is put in your junk folder?

Example

You've been keeping track of the last 1000 emails you received. You find that 100 of them are spam. You also find that 200 of them were put in your junk folder, of which 90 were spam.

What is the probability an email you receive is spam?

$$P(X) = 100/1000 = .1$$

What is the probability an email you receive is put in your junk folder?

$$P(Y) = 200/1000 = .2$$

Example

You've been keeping track of the last 1000 emails you received. You find that 100 of them are spam. You also find that 200 of them were put in your junk folder, of which 90 were spam.

What is the probability an email you receive is spam?

$$P(X) = 100/1000 = .1$$

What is the probability an email you receive is put in your junk folder?

$$P(Y) = 200/1000 = .2$$

Given that an email is in your junk folder, what is the probability it is spam?

Example

You've been keeping track of the last 1000 emails you received. You find that 100 of them are spam. You also find that 200 of them were put in your junk folder, of which 90 were spam.

What is the probability an email you receive is spam?

$$P(X) = 100/1000 = .1$$

What is the probability an email you receive is put in your junk folder?

$$P(Y) = 200/1000 = .2$$

Given that an email is in your junk folder, what is the probability it is spam?

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)} = .09/.2 = .45$$

Example

You've been keeping track of the last 1000 emails you received. You find that 100 of them are spam. You also find that 200 of them were put in your junk folder, of which 90 were spam.

What is the probability an email you receive is spam?

$$P(X) = 100/1000 = .1$$

What is the probability an email you receive is put in your junk folder?

$$P(Y) = 200/1000 = .2$$

Given that an email is in your junk folder, what is the probability it is spam?

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} = .09/.2 = .45$$

Given that an email is spam, what is the probability it is in your junk folder?

Example

You've been keeping track of the last 1000 emails you received. You find that 100 of them are spam. You also find that 200 of them were put in your junk folder, of which 90 were spam.

What is the probability an email you receive is spam?

$$P(X) = 100 / 1000 = .1$$

What is the probability an email you receive is put in your junk folder?

$$P(Y) = 200 / 1000 = .2$$

Given that an email is in your junk folder, what is the probability it is spam?

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)} = .09 / .2 = .45$$

Given that an email is spam, what is the probability it is in your junk folder?

$$P(Y \mid X) = \frac{P(X \cap Y)}{P(X)} = .09 / .1 = .9$$

Deriving Bayes Rule

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}$$

$$P(Y \mid X) = \frac{P(X \cap Y)}{P(X)}$$

Bayes rule :

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

Bayesian Learning

Application to Machine Learning

- In machine learning we have a space H of hypotheses:
  $h_1, \; h_2, \ldots, h_n$ (possibly infinite)

- We also have a set D of data

- We want to calculate $P(h \mid D)$

- Bayes rule gives us:

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

Terminolog y

- ***Prior probability* of *h*:**
  - *P(h)*: Probability that hypothesis *h* is true given our prior knowledge
  - If no prior knowledge, all $h \in H$ are equally probable
- ***Posterior probability* of *h*:**
  - *P(h | D)*: Probability that hypothesis *h* is true, given the data *D*.
- ***Likelihood* of *D*:**
  - *P(D | h)*: Probability that we will see data *D*, given hypothesis *h* is true.
- ***Marginal likelihood of D***
  - $$P(D) = \sum_h P(D \mid h) P(h)$$

The Monty Hall Problem

You are a contestant on a game show.

There are 3 doors, A, B, and C. There is a new car behind one of them and goats behind the other two.

Monty Hall, the host, knows what is behind the doors. He asks you to pick a door, any door. You pick door A.

Monty tells you he will open a door, different from A, that has a goat behind it. He opens door B: behind it there is a goat.

Monty now gives you a choice: Stick with your original choice A or switch to C.

Should you switch?

http://math.ucsd.edu/~crypto/Monty/monty.html

Bayesian probability formulation

Hypothesis space *H*:

h_1 = Car is behind door A
h_2 = Car is behind door B
h_3 = Car is behind door C

Data *D*: After you picked door A, Monty opened B to show a goat

What is $P(h_1 \mid D)$?
What is $P(h_2 \mid D)$?
What is $P(h_3 \mid D)$?

Prior probability:

$P(h_1) = 1/3$ $P(h_2) = 1/3$ $P(h_3) = 1/3$

Likelihood:

$P(D \mid h_1) = 1/2$
$P(D \mid h_2) = 0$
$P(D \mid h_3) = 1$

Marginal likelihood:

$$P(D) = p(D|h_1)p(h_1) + p(D|h_2)p(h_2) + p(D|h_3)p(h_3) = 1/6 + 0 + 1/3 = 1/2$$

By Bayes rule:

$$P(h_1 \mid D) = \frac{P(D \mid h_1)P(h_1)}{P(D)} = \left(\frac{1}{2}\right)\left(\frac{1}{3}\right)(2) = \frac{1}{3}$$

$$P(h_2 \mid D) = \frac{P(D \mid h_2)P(h_2)}{P(D)} = (0)\left(\frac{1}{3}\right)(2) = 0$$

$$P(h_3 \mid D) = \frac{P(D \mid h_3)P(h_3)}{P(D)} = (1)\left(\frac{1}{3}\right)(2) = \frac{2}{3}$$

So you should switch!

MAP ("maximum a posteriori") Learning

Bayes rule: $$P(h \mid D) = \frac{P(D \mid h) P(h)}{P(D)}$$

Goal of learning: Find maximum a posteriori hypothesis h_{MAP}:

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} \, P(h \mid D)$$

$$= \underset{h \in H}{\operatorname{argmax}} \frac{P(D \mid h) P(h)}{P(D)}$$

$$= \underset{h \in H}{\operatorname{argmax}} \, P(D \mid h) P(h)$$

because $P(D)$ is a constant independent of h.

Note: If every $h \in H$ is equally probable, then

$$h_{\text{MAP}} = \operatorname*{argmax}_{h \in H} P(D \mid h)$$

h_{MAP} is called the "maximum likelihood hypothesis".

A Medical Example

Toby takes a test for leukemia. The test has two outcomes: positive and negative. It is known that if the patient has leukemia, the test is positive 98% of the time. If the patient does not have leukemia, the test is positive 3% of the time. It is also known that 0.008 of the population has leukemia.

Toby's test is positive.

Which is more likely: Toby has leukemia or Toby does not have leukemia?

- **Hypothesis space:**

  h_1 = T. has leukemia

  h_2 = T. does not have leukemia

- **Prior:** 0.008 of the population has leukemia. Thus

  $P(h_1) = 0.008$

  $P(h_2) = 0.992$

- **Likelihood:**

  $P(+ \mid h_1) = 0.98,\ P(- \mid h_1) = 0.02$

  $P(+ \mid h_2) = 0.03,\ P(- \mid h_2) = 0.97$

- **Posterior knowledge:**

  Blood test is + for this patient.

- In summary

$$P(h_1) = 0.008,\ P(h_2) = 0.992$$

$$P(+ \mid h_1) = 0.98,\ P(- \mid h_1) = 0.02$$

$$P(+ \mid h_2) = 0.03,\ P(- \mid h_2) = 0.97$$

- Thus:

$$h_{MAP} = \underset{h \in H}{argmax}\ P(D \mid h) P(h)$$

$$P(+ \mid leukemia) P(leukemia) = (0.98)(0.008) = 0.0078$$

$$P(+ \mid \neg leukemia) P(\neg leukemia) = (0.03)(0.992) = 0.0298$$

$$h_{MAP} \boxed{= \neg leukemia}$$

- What is P(leukemia|+)?

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

So,

$$P(leukemia \mid +) = \frac{0.0078}{0.0078 + 0.0298} = 0.21$$

$$P(\neg leukemia \mid +) = \frac{0.0298}{0.0078 + 0.0298} = 0.79$$

These are called the “posterior” probabilities.

In-Class Exercises, Part 1

Bayesianism vs. Frequentism

- Classical probability: **Frequentists**
  - Probability of a particular event is defined relative to its *frequency* in a sample space of events.

  - E.g., probability of "the coin will come up heads on the next trial" is defined relative to the *frequency* of heads in a sample space of coin tosses.

- **Bayesian** probability:
  - Combine measure of "prior" belief you have in a proposition with your subsequent observations of events.

- **Example:** Bayesian can assign probability to statement "There was life on Mars a billion years ago" but frequentist cannot.

Independence and Conditional Independence

- Two random variables, X and Y, are independent if

$$P(X,Y) = P(X)P(Y)$$

- Two random variables, X and Y, are independent *given* C if

$$P(X,Y \mid C) = P(X \mid C)P(Y \mid C)$$

Naive Bayes Classifier

Let $f(\mathbf{x})$ be a target function for classification: $f(\mathbf{x}) \in \{+1, -1\}$.

Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$

We want to find the most probable class value, h_{MAP}, given the data $\mathbf{x}$:

$$class_{MAP} = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} \; P(class \mid D)$$

$$= \underset{class \in \{+1,-1\}}{\operatorname{argmax}} \; P(class \mid x_1, x_2, \ldots, x_n)$$

By Bayes Theorem:

$$class_{MAP} = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} \frac{P(x_1, x_2, \ldots, x_n \mid class) P(class)}{P(x_1, x_2, \ldots, x_n)}$$

$$= \underset{class \in \{+1,-1\}}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid class) P(class)$$

$P(class)$ can be estimated from the training data. How?

However, in general, not practical to use training data to estimate $P(x_1, x_2, \ldots, x_n \mid class)$. Why not?

- Naive Bayes classifier: Assume

$$P(x_1, x_2, \ldots, x_n \mid class) = P(x_1 \mid class) P(x_2 \mid class) \cdots P(x_n \mid class)$$

Is this a good assumption?

Given this assumption, here's how to classify an instance $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$:

Naive Bayes classifier:

$$class_{NB}(\mathbf{x}) = \operatorname*{argmax}_{class \in \{+1,-1\}} P(class) \prod_i P(x_i \mid class)$$

- Naive Bayes classifier: Assume

$$P(x_1, x_2, \ldots, x_n \mid class) = P(x_1 \mid class)P(x_2 \mid class) \cdots P(x_n \mid class)$$

Is this a good assumption?

Given this assumption, here's how to classify an instance $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$:

Naive Bayes classifier:

$$\boxed{class_{NB}(\mathbf{x}) = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} \; P(class)\prod_i P(x_i \mid class)}$$

- Naive Bayes classifier: Assume

$$P(x_1, x_2, \ldots, x_n \mid class) = P(x_1 \mid class) P(x_2 \mid class) \cdots P(x_n \mid class)$$

Is this a good assumption?

Given this assumption, here's how to classify an instance $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$:

Naive Bayes classifier:

$$\boxed{class_{NB}(\mathbf{x}) = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} \; P(class) \prod_i P(x_i \mid class)}$$

To train: Estimate the values of these various probabilities over the training set.

Training data:

Day	Outlook	Temp	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

Test data:

D15	Sunny	Cool	High	Strong	?

Use training data to compute a probabilistic *model*:

$$P(Outlook=Sunny\,|\,Yes)=2/9 \quad P(Outlook=Sunny\,|\,No)=3/5$$

$$P(Outlook=Overcast\,|\,Yes)=4/9 \quad P(Outlook=Overcast\,|\,No)=0$$

$$P(Outlook=Rain\,|\,Yes)=3/9 \quad P(Outlook=Rain\,|\,No)=2/5$$

$$P(Temperature=Hot\,|\,Yes)=2/9 \quad P(Temperature=Hot\,|\,No)=2/5$$

$$P(Temperature=Mild\,|\,Yes)=4/9 \quad P(Temperature=Mild\,|\,No)=2/5$$

$$P(Temperature=Cool\,|\,Yes)=3/9 \quad P(Temperature=Cool\,|\,No)=1/5$$

$$P(Humidity=High\,|\,Yes)=3/9 \quad P(Humidity=High\,|\,No)=4/5$$

$$P(Humidity=Normal\,|\,Yes)=6/9 \quad P(Humidity=Normal\,|\,No)=1/5$$

$$P(Wind=Strong\,|\,Yes)=3/9 \quad P(Wind=Strong\,|\,No)=3/5$$

$$P(Wind=Weak\,|\,Yes)=6/9 \quad P(Wind=Weak\,|\,No)=2/5$$

Use training data to compute a probabilistic *model*:

$$P(Outlook = Sunny \mid Yes) = 2/9 \quad P(Outlook = Sunny \mid No) = 3/5$$
$$P(Outlook = Overcast \mid Yes) = 4/9 \quad P(Outlook = Overcast \mid No) = 0$$
$$P(Outlook = Rain \mid Yes) = 3/9 \quad P(Outlook = Rain \mid No) = 2/5$$

$$P(Temperature = Hot \mid Yes) = 2/9 \quad P(Temperature = Hot \mid No) = 2/5$$
$$P(Temperature = Mild \mid Yes) = 4/9 \quad P(Temperature = Mild \mid No) = 2/5$$
$$P(Temperature = Cool \mid Yes) = 3/9 \quad P(Temperature = Cool \mid No) = 1/5$$

$$P(Humidity = High \mid Yes) = 3/9 \quad P(Humidity = High \mid No) = 4/5$$
$$P(Humidity = Normal \mid Yes) = 6/9 \quad P(Humidity = Normal \mid No) = 1/5$$

$$P(Wind = Strong \mid Yes) = 3/9 \quad P(Wind = Strong \mid No) = 3/5$$
$$P(Wind = Weak \mid Yes) = 6/9 \quad P(Wind = Weak \mid No) = 2/5$$

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D15	Sunny	Cool	High	Strong	?

Use training data to compute a probabilistic *model*:

$$P(Outlook = Sunny \mid Yes) = 2/9 \quad P(Outlook = Sunny \mid No) = 3/5$$
$$P(Outlook = Overcast \mid Yes) = 4/9 \quad P(Outlook = Overcast \mid No) = 0$$
$$P(Outlook = Rain \mid Yes) = 3/9 \quad P(Outlook = Rain \mid No) = 2/5$$

$$P(Temperature = Hot \mid Yes) = 2/9 \quad P(Temperature = Hot \mid No) = 2/5$$
$$P(Temperature = Mild \mid Yes) = 4/9 \quad P(Temperature = Mild \mid No) = 2/5$$
$$P(Temperature = Cool \mid Yes) = 3/9 \quad P(Temperature = Cool \mid No) = 1/5$$

$$P(Humidity = High \mid Yes) = 3/9 \quad P(Humidity = High \mid No) = 4/5$$
$$P(Humidity = Normal \mid Yes) = 6/9 \quad P(Humidity = Normal \mid No) = 1/5$$

$$P(Wind = Strong \mid Yes) = 3/9 \quad P(Wind = Strong \mid No) = 3/5$$
$$P(Wind = Weak \mid Yes) = 6/9 \quad P(Wind = Weak \mid No) = 2/5$$

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D15	Sunny	Cool	High	Strong	?

$$\boxed{class_{NB}(\mathbf{x}) = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} P(class) \prod_i P(x_i \mid class)}$$

Exercises, Part 2, #1

Estimating probabilities / Smoothing

- **Recap:** In previous example, we had a training set and a new example,

  (Outlook=sunny, Temperature=cool, Humidity=high, Wind=strong)

- We asked: What classification is given by a naive Bayes classifier?

- Let n_c be the number of training instances with class c.

  E.g., $n_{yes} = 9$

- Let $n_c^{x_i=a_k}$ be the number of training instances with attribute value $x_i=a_k$ and class c.

  E.g., $n_{yes}^{outlook=sunny} = 2$

Then $P(x_i = a_i \mid c) = \dfrac{n_c^{x_i=a_k}}{n_c}$

E.g., $P(outlook = sunny \mid yes) = \dfrac{2}{9}$

- **Problem with this method:** If n_c is very small, gives a poor estimate.

- E.g., $P(Outlook = Overcast \mid no) = 0$.

- Now suppose we want to classify a new instance:

(Outlook=overcast, Temperature=cool, Humidity=high, Wind=strong)

Then:

$$P(\text{no})\prod_i P(x_i \mid \text{no}) = 0$$

This incorrectly gives us zero probability due to small sample.

Training data:

Day	Outlook	Temp	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

How should we modify probabilities?

One solution: Laplace smoothing (also called "add-one" smoothing)

For each class c and attribute x_i with value a_k, add one "virtual" instance.

That is, for each class c, recalculate:

$$P(x_i = a_i \mid c) = \frac{n_c^{x_i = a_k} + 1}{n_c + K}$$

where K is the number of possible values of attribute a.

Training data:

Day	Outlook	Temp	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

Laplace smoothing: Add the following virtual instances for *Outlook*:

Outlook=Sunny: **Yes** *Outlook=Overcast*: **Yes** *Outlook=Rain*: **Yes**

Outlook=Sunny: **No** *Outlook=Overcast*: **No** *Outlook=Rain*: **No**

$$P(Outlook = overcast \mid \mathbf{No}) = \frac{0}{5} \quad \rightarrow \quad \frac{n_c^{x_i=a_k} + 1}{n_c + K} = \frac{0+1}{5+3} = \frac{1}{8}$$

$$P(Outlook = overcast \mid \mathbf{Yes}) = \frac{4}{9} \quad \rightarrow \quad \frac{n_c^{x_i=a_k} + 1}{n_c + K} = \frac{4+1}{9+3} = \frac{5}{12}$$

$$P(Outlook = Sunny \mid Yes) = 2/9 \rightarrow 3/12 \quad P(Outlook = Sunny \mid No) = 3/5 \rightarrow 4/8$$
$$P(Outlook = Overcast \mid Yes) = 4/9 \rightarrow 5/12 \quad P(Outlook = Overcast \mid No) = 0/5 \rightarrow 1/8$$
$$P(Outlook = Rain \mid Yes) = 3/9 \rightarrow 4/12 \quad P(Outlook = Rain \mid No) = 2/5 \rightarrow 3/8$$

$$P(Humidity = High \mid Yes) = 3/9 \rightarrow 4/11 \quad P(Humidity = High \mid No) = 4/5 \rightarrow 5/7$$
$$P(Humidity = Normal \mid Yes) = 6/9 \rightarrow 7/11 \quad P(Humidity = Normal \mid No) = 1/5 \rightarrow 2/7$$

Etc.

Exercises, part 2, #2

Naive Bayes on continuous-valued attributes

- How to deal with continuous-valued attributes?

Two possible solutions:

- Discretize

- Assume particular probability distribution of classes over values (estimate parameters from training data)

Discretization: Equal-Width Binning

For each attribute x_i, create k equal-width bins in interval from $min(x_i)$ to $max(x_i)$.

The discrete "attribute values" are now the bins.

Questions: What should k be? What if some bins have very few instances?

Problem with balance between *discretization bias* and *variance*.

The more bins, the lower the bias, but the higher the variance, due to small sample size.

Discretization: Equal-Frequency Binning

For each attribute x_i , create k bins so that each bin contains an equal number of values.

Also has problems: What should k be? Hides outliers. Can group together instances that are far apart.

Gaussian Naïve Bayes

Assume that within each class, values of each numeric feature are normally distributed:

$$p(x_i \mid c) = N(x_i; \mu_{i,c}, \sigma_{i,c})$$

where

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

where $\mu_{i,c}$ is the mean of feature *i* given the class c, and $\sigma_{i,c}$ is the standard deviation of feature *i* given the class c

We estimate $\mu_{i,c}$ and $\sigma_{i,c}$ from training data.

Example

x_1	x_2	**Class**
3.0	5.1	**POS**
4.1	6.3	**POS**
7.2	9.8	**POS**
2.0	1.1	NEG
4.1	2.0	NEG
8.1	9.4	NEG

Example

x_1	x_2	**Class**
3.0	5.1	**POS**
4.1	6.3	**POS**
7.2	9.8	**POS**
2.0	1.1	NEG
4.1	2.0	NEG
8.1	9.4	NEG

$$P(\textbf{POS}) = 0.5$$

$$P(\textbf{NEG}) = 0.5$$

$$\mu_{1,\text{POS}} = \frac{(3.0 + 4.1 + 7.2)}{3} = 4.8$$

$$\sigma_{1,\text{POS}} = \sqrt{\frac{(3.0 - 4.8)^2 + (4.1 - 4.8)^2 + (7.2 - 4.8)^2}{3}} = 1.8$$

$$\mu_{1,\text{NEG}} = \frac{(2.0 + 4.1 + 8.1)}{3} = 4.7$$

$$\sigma_{1,\text{NEG}} = \sqrt{\frac{(2.0 - 4.7)^2 + (4.1 - 4.7)^2 + (8.1 - 4.7)^2}{3}} = 2.5$$

$$\mu_{2,\text{POS}} = \frac{(5.1 + 6.3 + 9.8)}{3} = 7.1$$

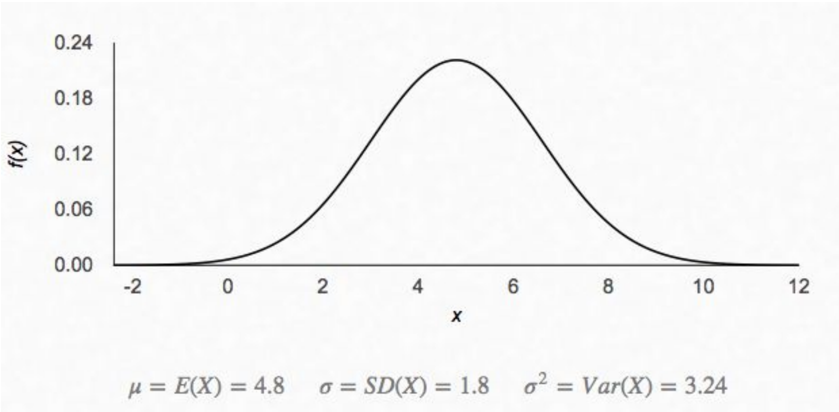
$$\sigma_{2,\text{POS}} = \sqrt{\frac{(5.1 - 7.1)^2 + (6.3 - 7.1)^2 + (9.8 - 7.1)^2}{3}} = 2.0$$

$$\mu_{2,\text{NEG}} = \frac{(1.1 + 2.0 + 9.4)}{3} = 4.2$$

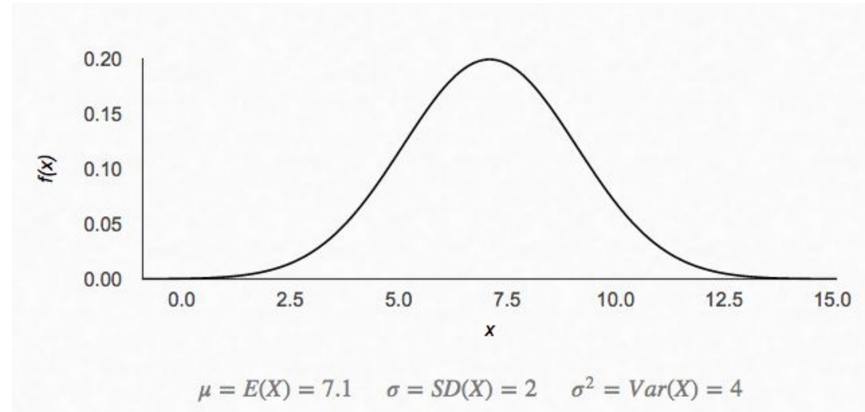
$$\sigma_{2,\text{NEG}} = \sqrt{\frac{(1.1 - 4.2)^2 + (2.0 - 4.2)^2 + (9.4 - 4.2)^2}{3}} = 3.7$$

http://homepage.stat.uiowa.edu/~mbognar/applets/normal.html

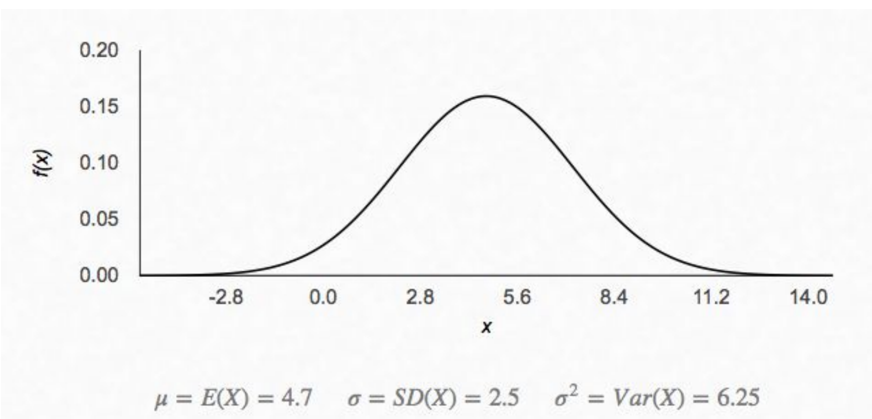
$N_{1,\mathbf{POS}} = N(x;\, 4.8,\, 1.8)$



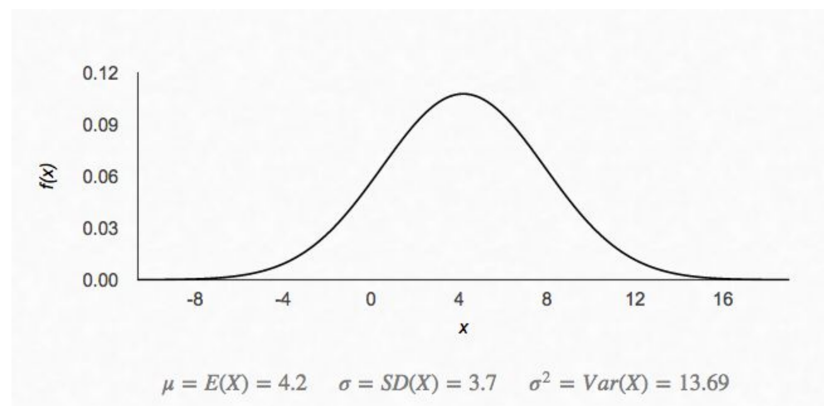
$N_{2,\mathbf{POS}} = N(x;\, 7.1,\, 2.0)$



$N_{1,\mathbf{NEG}} = N(x;\, 4.7,\, 2.5)$



$N_{2,\mathbf{NEG}} = N(x;\, 4.2,\, 3.7)$



Now, suppose you have a new example **x**, with $x_1 = 5.2, x_2 = 6.3$.

What is $class_{NB}(\mathbf{x})$?

Now, suppose you have a new example **x**, with $x_1 = 5.2, x_2 = 6.3$.

What is $class_{NB}(\mathbf{x})$?

$$class_{NB}(\mathbf{x}) = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} \; P(class) \prod_i P(x_i \mid class)$$

$$P(x_i \mid c) = N(x_i; \mu_{i,c}, \sigma_{i,c})$$

where

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

Note: N is the probability density function, but can be used analogously to probability in Naïve Bayes calculations.

Now, suppose you have a new example **x**, with $x_1 = 5.2$, $x_2 = 6.3$.

What is $class_{NB}(\mathbf{x})$?

$$class_{NB}(\mathbf{x}) = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} \; P(class) \prod_i P(x_i \mid class)$$

$$P(x_i \mid c) = N(x_i; \boldsymbol{\mu}_{i,c}, \boldsymbol{\sigma}_{i,c})$$

where

$$N(x; \boldsymbol{\mu}, \boldsymbol{\sigma}) = \frac{1}{\sqrt{2\pi}\boldsymbol{\sigma}} e^{-\frac{(x-\boldsymbol{\mu})^2}{2\sigma^2}}$$

$$P(x_1 \mid \mathbf{POS}) = \frac{1}{\sqrt{2\pi}(1.8)} e^{-\frac{(5.2-4.8)^2}{2(1.8)^2}} = .22$$

$$P(x_2 \mid \mathbf{POS}) = \frac{1}{\sqrt{2\pi}(2.0)} e^{-\frac{(6.3-7.1)^2}{2(2.0)^2}} = .18$$

$$P(x_1 \mid \mathbf{NEG}) = \frac{1}{\sqrt{2\pi}(2.5)} e^{-\frac{(5.2-4.7)^2}{2(2.5)^2}} = .16$$

$$P(x_2 \mid \mathbf{NEG}) = \frac{1}{\sqrt{2\pi}(3.7)} e^{-\frac{(6.3-4.2)^2}{2(3.7)^2}} = .09$$

$Positive:$

$$P(\mathbf{POS})P(x_1 \mid \mathbf{POS})P(x_2 \mid \mathbf{POS}) = (.5)(.22)(.18) = .02$$

$Negative:$

$$P(\mathbf{NEG})P(x_1 \mid \mathbf{NEG})P(x_2 \mid \mathbf{NEG}) = (.5)(.16)(.09) = .0072$$

$$class_{NB}(\mathbf{x}) = \mathbf{POS}$$

Use logarithms to avoid underflow

$$class_{NB}(\mathbf{x}) = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} \; P(class) \prod_i P(x_i \mid class)$$

$$= \underset{class \in \{+1,-1\}}{\operatorname{argmax}} \; \log\left(P(class) \prod_i P(x_i \mid class) \right)$$

$$= \underset{class \in \{+1,-1\}}{\operatorname{argmax}} \left(\log P(class) + \sum_i \log P(x_i \mid class) \right)$$