CS4710: Artificial Intelligence
Intro to Machine Learning

Let's look at a couple more machine learning algorithms!

Topics

- Another learning algorithm
- Naïve Bayes Classifier
  - More Bayesian Networks!!

Naïve-Bayes Classifier

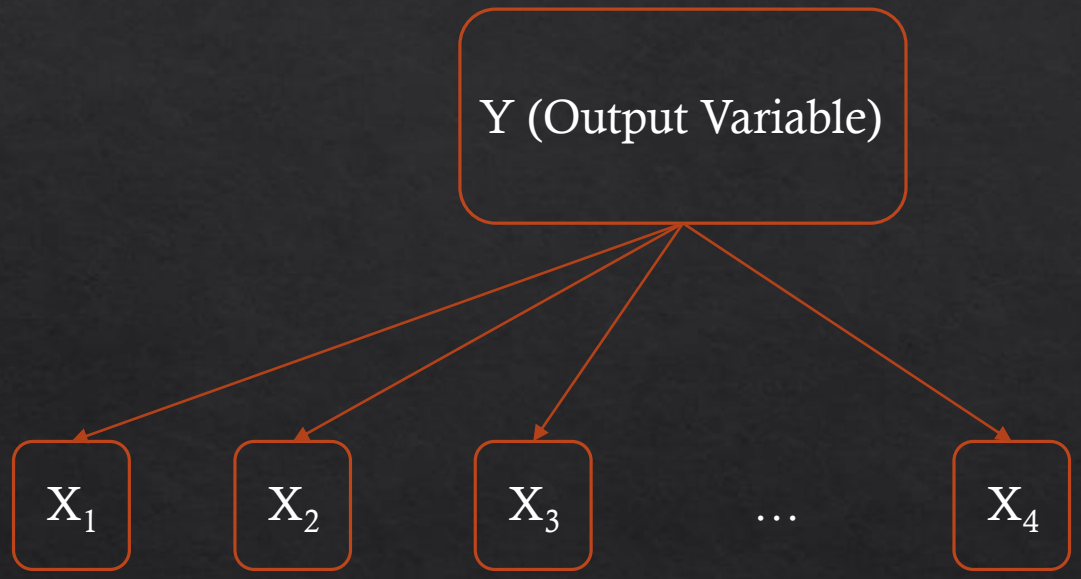
Naïve-Bayes Classifier

- Another Classification Algorithm
- Main Idea: Use a simple Bayesian Network (they're back!) to do classification
- Main Issues:
  - Need to build network from training set
  - A few important assumptions about features
  - Algorithm slightly different for discrete and continuous features

Naïve-Bayes Classifier

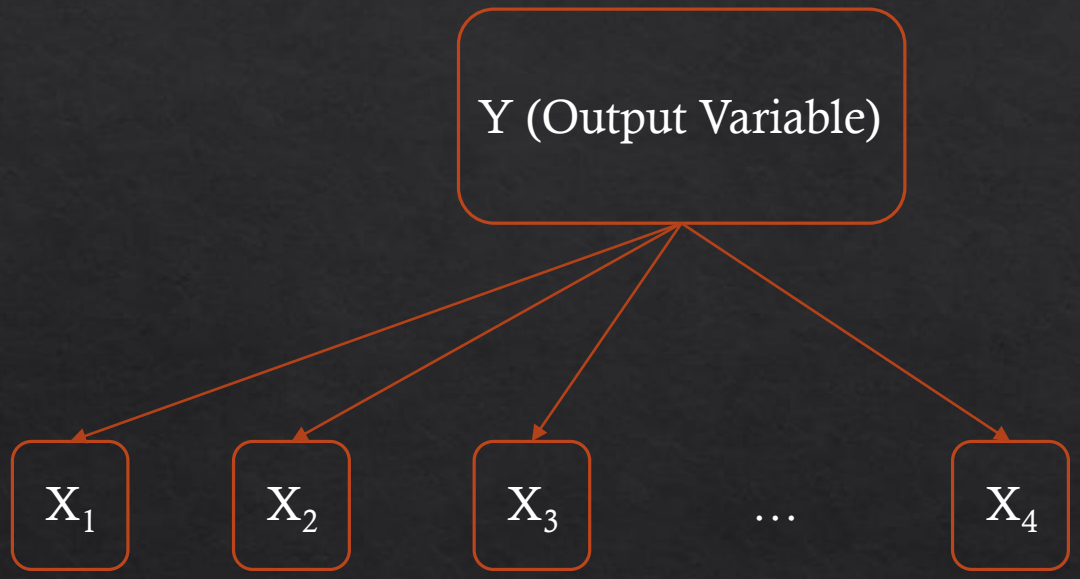
- Training set ***x*** contains ***n features***
  - As before, these features can be anything but should be numeric
- Output variable is one of two classifications
  - $y \in \{0, 1\}$
  - 0 is one class (e.g., email not spam)
  - 1 is other class (e.g., email is spam)
- Naïve Bayes extends very easily to larger number of classes

Naïve-Bayes Classifier



- Treats learning as a probability estimation problem.
- IDEA!: Set up the Bayesian Network as seen here
- Our goal then is to predict:
  - $P(Y = y_i \mid X = \{x_1, x_2, \ldots\})$
  - Y is the output variable
  - X is the set of evidence we are given by the example data row

Bayes-Rule



- Bayes' Rule States That:

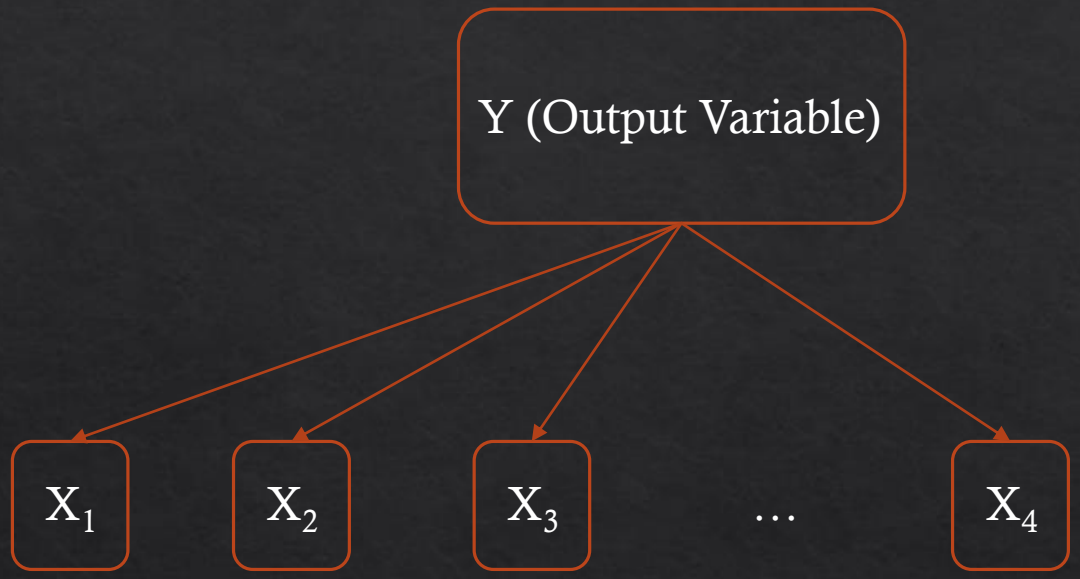
$$P(Y = yj|X = xk) = \frac{P(X = xk|Y = yj)P(Y = yj)}{P(X = xk)}$$

- Remember that :

$$P(X = xk) = \sum_{j} P(X = xk|Y = yj)P(Y = yj)$$

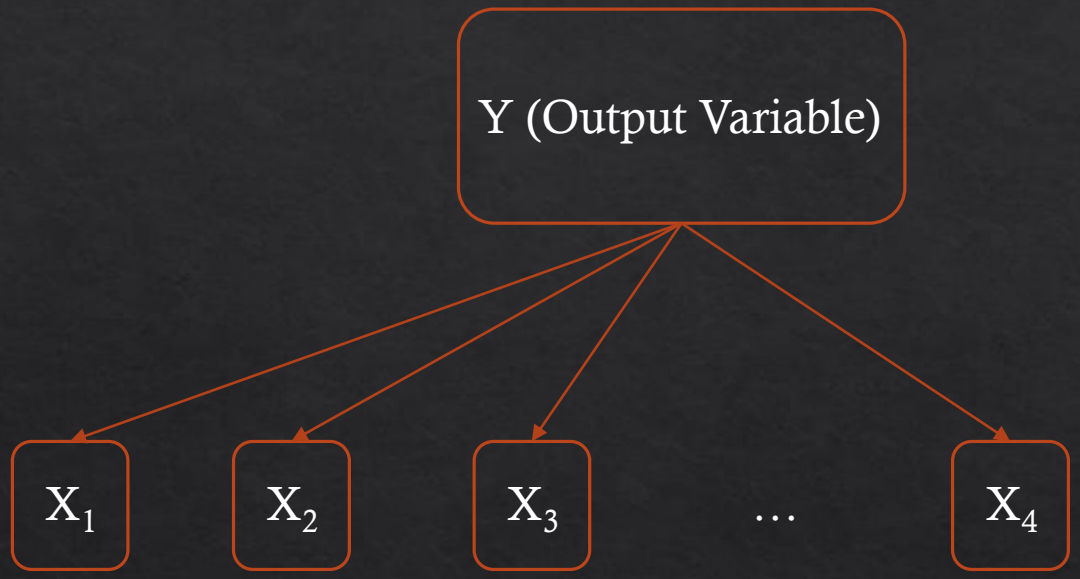
- Instead, of estimating P(Y | X) directly, we:
  - Estimate P(Y) and P(X | Y) from training set
  - Use the formula above to compute P(Y | X) for new examples

Scalability

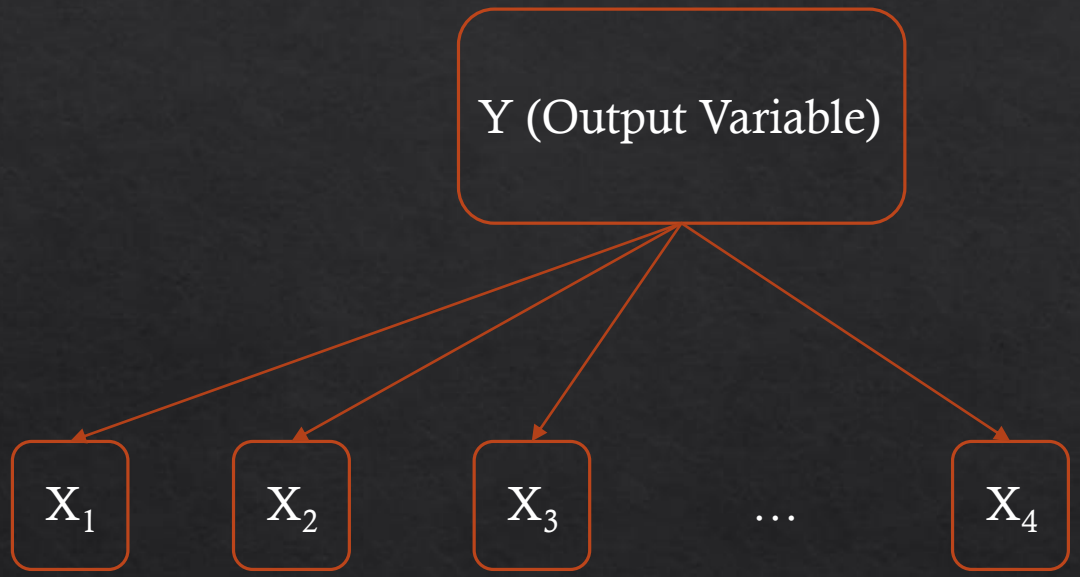


- How many of these P(X | Y) formulas are there to estimate?
  - If every variable can take two values, then we only estimate $\theta(n)$ total values
- If we were trying to estimate using an unbiased bayesian classifier, then we would have to estimate every combo:
  - $\theta_{ij} = P(Y = yi|X = xj)$
  - There are 2^n combos of these…way too many combinations of evidence variables

Scalability



- How many of these P(X | Y) formulas are there to estimate?
  - If every variable can take two values, then we only estimate $\theta(2n)$ total values
- ^^^ We can ONLY do this if our features are conditionally independent of one another. This is a fairly large assumption and you must take care to ensure your dataset follows this assumption.
- Conditional Independence:
  - P(X | Y,Z) = P(X | Z) means that Y and Z are C.I.



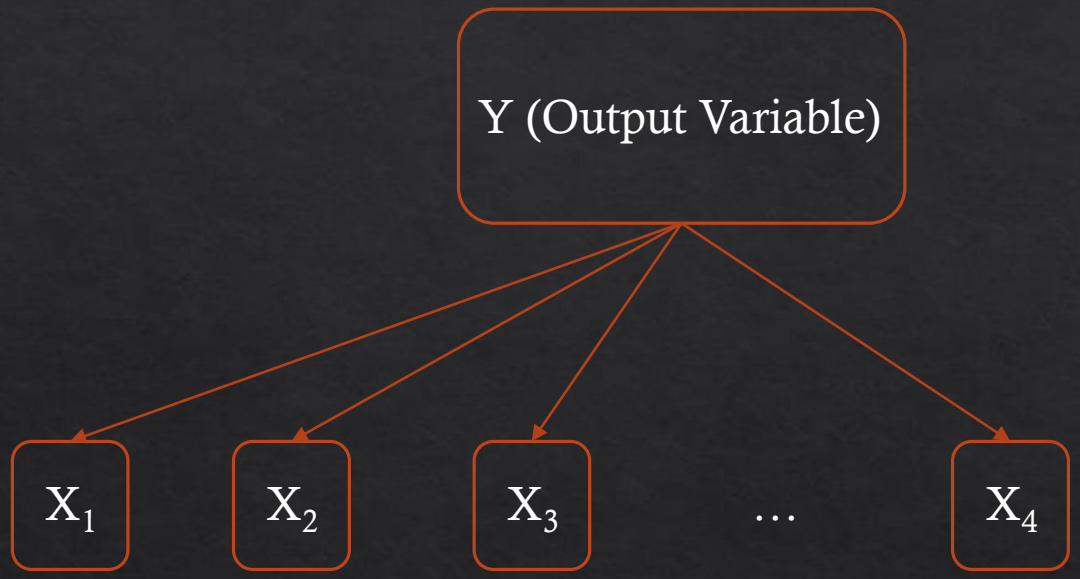
Summary So Far

- 1. Build the Bayesian network seen here
- 2. Learn the probabilities P(X | Y) and P(Y) from the training data
  - We'll see exactly how in a second
- 3. When given a new row of test data, compute:

$$P(Y = yk|X_1, \ldots Xn) = \frac{P(Y = yk)\prod_i P(Xi|Y = yk)}{\sum_j P(Y = yj)\prod_i P(Xi|Y = yj)}$$

And output the following classification:

$$Y \leftarrow argMax_{y_k} \frac{P(Y = yk)\prod_i P(X_i|Y = yk)}{\sum_j P(Y = yj)\prod_i P(Xi|Y = yj)}$$



Summary So Far

And output the following classification:

$$Y \leftarrow argMax_{y_k} \frac{P(Y = yk) \prod_i P(X_i|Y = yk)}{\sum_j P(Y = yj) \prod_i P(Xi|Y = yj)}$$

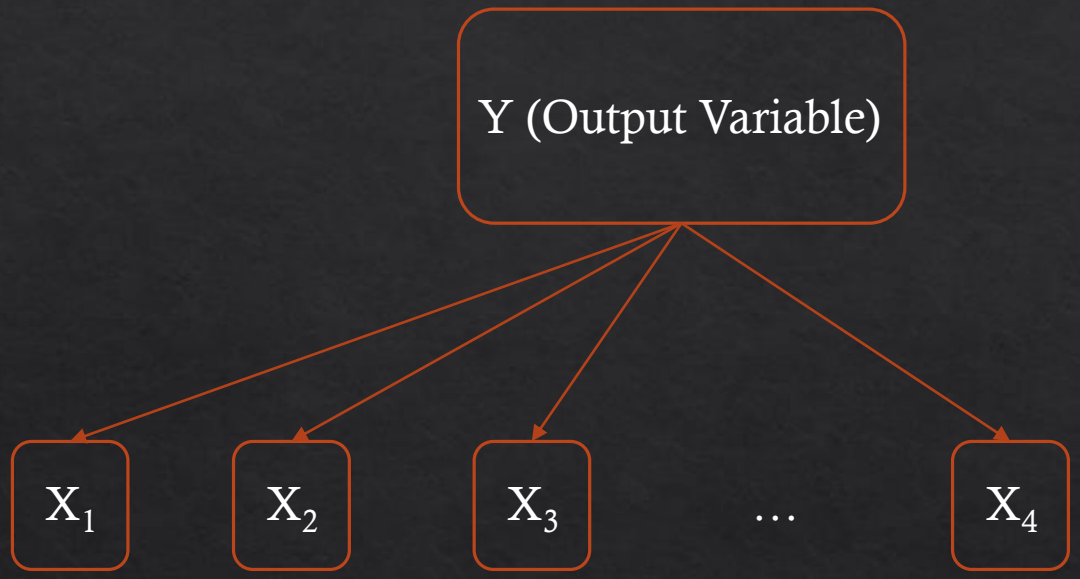
Notice the denominator does not depend on y_k so we can simplify to:

$$Y \leftarrow argMax_{y_k} P(Y = yk) \prod_i P(X_i|Y = yk)$$

Naïve-Bayes Classifier

Learning the network probabilities: Discrete valued data

Summary So Far



The last piece is to show how to use the training data to learn the probabilities of interest:

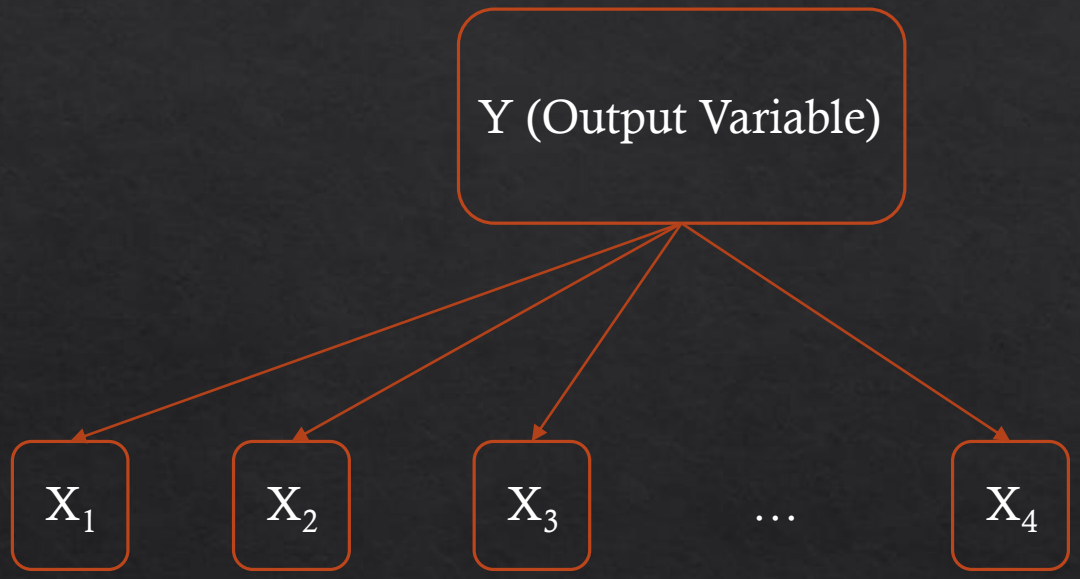
$P(Y)$

$P(X \mid Y)$

When all of the data is discrete, this is quite simple.

Remember, discrete here means that each feature and output (Y) have a finite number of values they can take on.

Estimating P(Y)



First let's learn P(Y) from the data. The prime indicates that these are estimates (not the actual values)

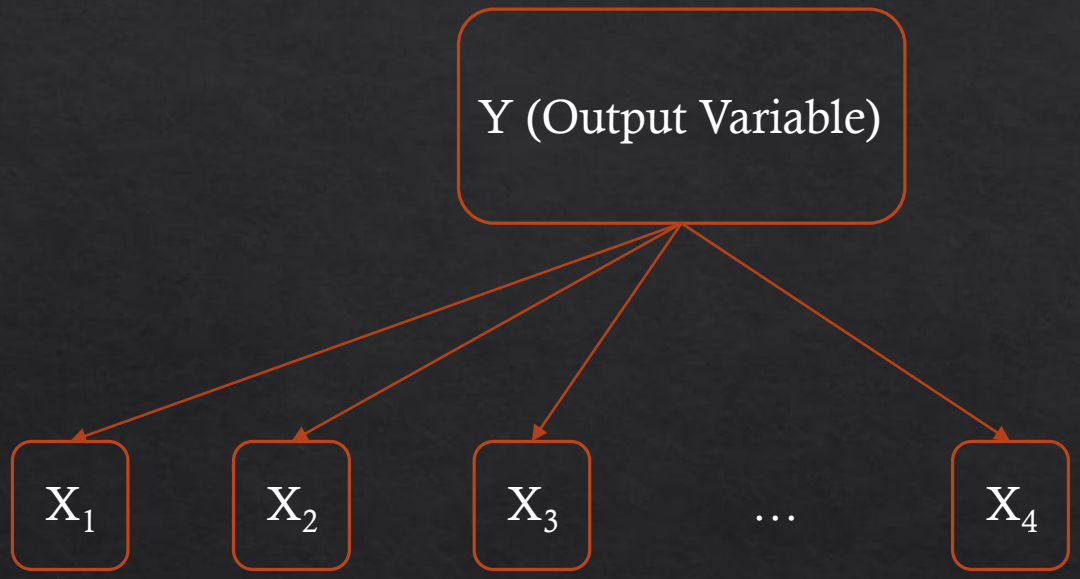
Given a training set, simply compute the following value:

$$\pi'_k = P'(Y = yk) = \frac{\#D\{Y = yk\}}{|D|}$$

Here, D is our training data and #D is an operator that counts the number of rows given some requirement.

So, #D{Y = y_k} is the number of rows where the output is exactly y_k

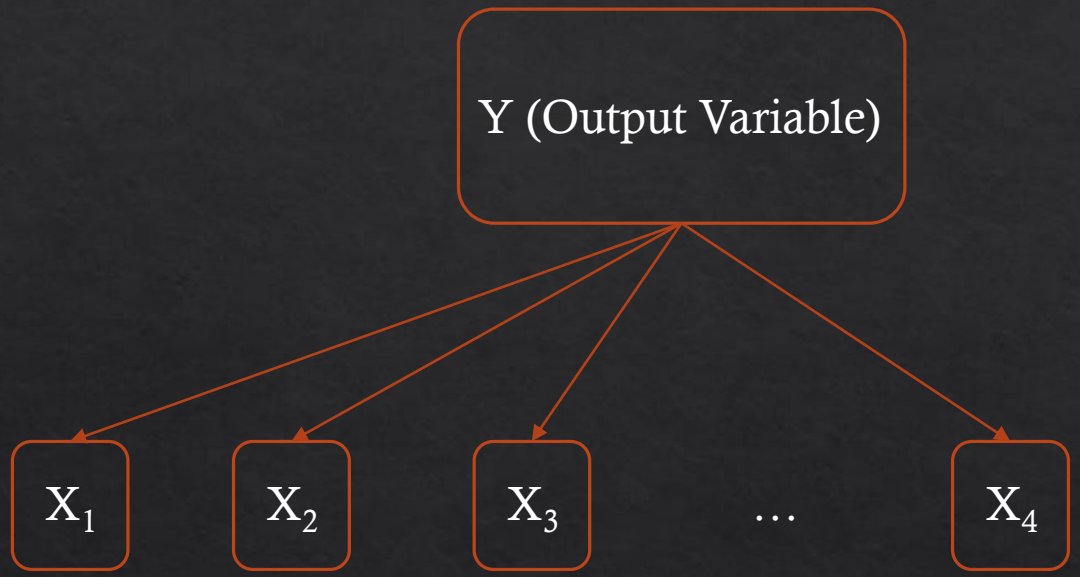
Estimating P(X | Y)



Now let's learn P(X | Y) from the data.

$$\theta'_{ijk} = P'(X_i = xij | Y = yk) = \frac{\#D\{Xi = xij \text{ ^ } Y = yk\}}{\#D\{Y = yk\}}$$

Estimating P(X | Y)



Now let's learn P(X | Y) from the data.

$$\theta'_{ijk} = P'(X_i = xij | Y = yk) = \frac{\#D\{Xi = xij \text{ ^ } Y = yk\}}{\#D\{Y = yk\}}$$

Sometimes, we add a dummy term:

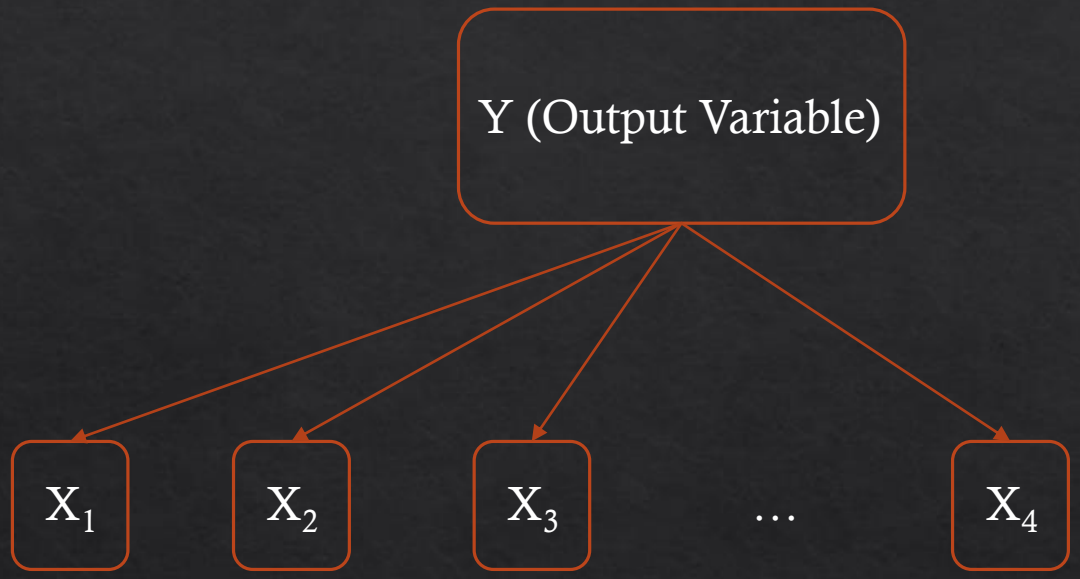
$$\theta'_{ijk} = P'(X_i = xij | Y = yk) = \frac{\#D\{X_i = xij \text{ ^ } Y = yk\} + l}{\#D\{Y = yk\} + lJ}$$

Where J is number of values X_i can take on and l is a "smoothing factor". Why might we need this?

Naïve-Bayes Classifier

Learning the network probabilities: Continuous Valued Data

Continuous Features



The last piece is to show how to use the training data to learn the probabilities of interest:

P(Y)

P(X | Y)

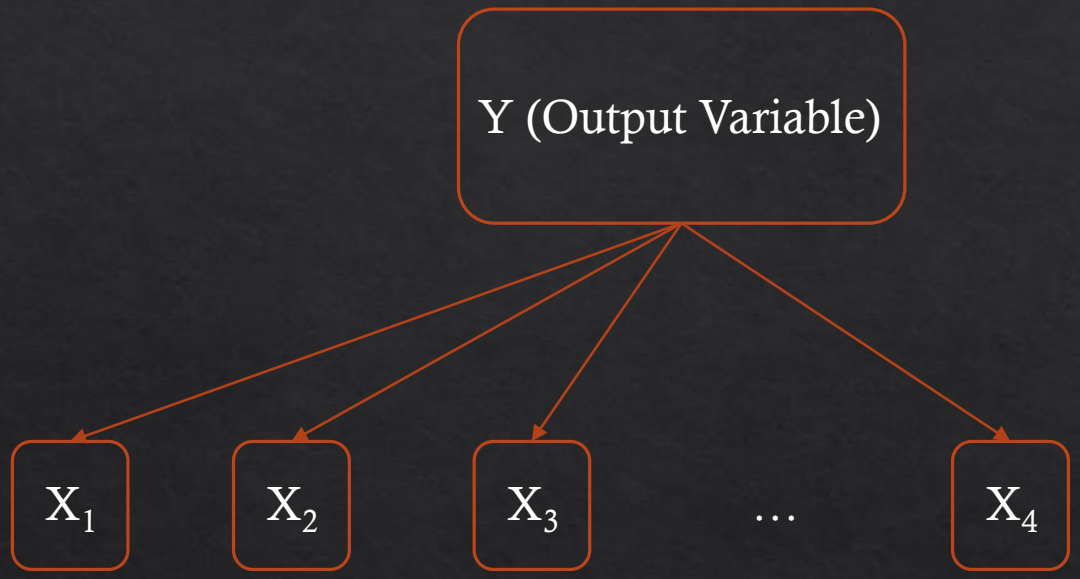
When features are ***continuous***, this is slightly more complicated.

Short version:

Instead let's estimate the mean and std. dev. of each

We can compute probability density function from this

Continuous Features



So, each feature X_i is assumed to have a Gaussian Distribution

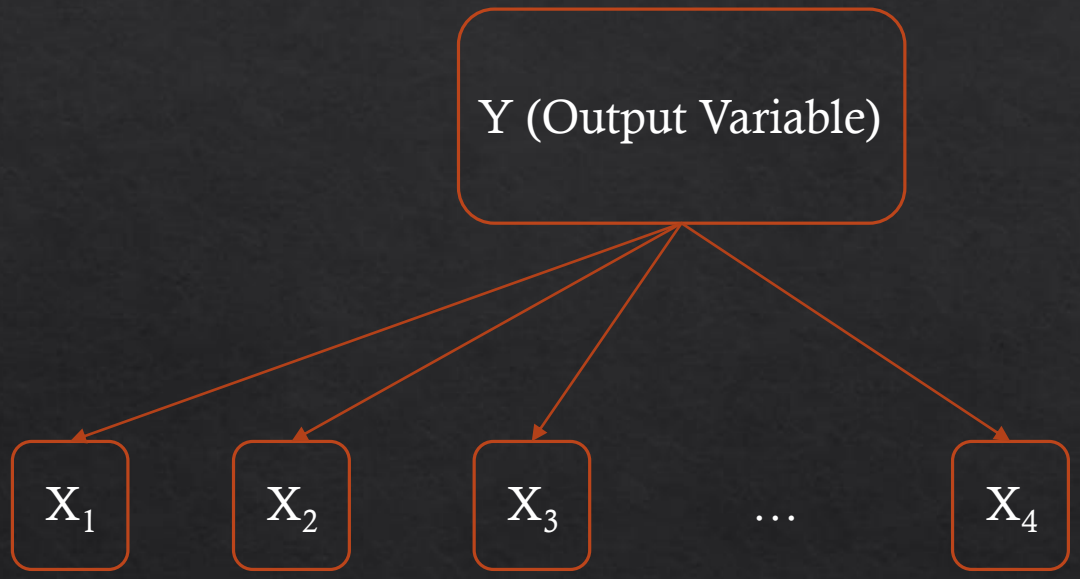
- i.e., normal distribution

We will thus estimate the following for each:

$$\mu_{ik} = E[X_i \mid Y = yk]$$
$$\sigma^2{}_{ik} = E[(X_i - \mu_{ik})^2 \mid Y = yk]$$

How many of these means and variances are there total that we need to estimate?

Continuous Features



So, each feature Xi is assumed to have a Gaussian Distribution

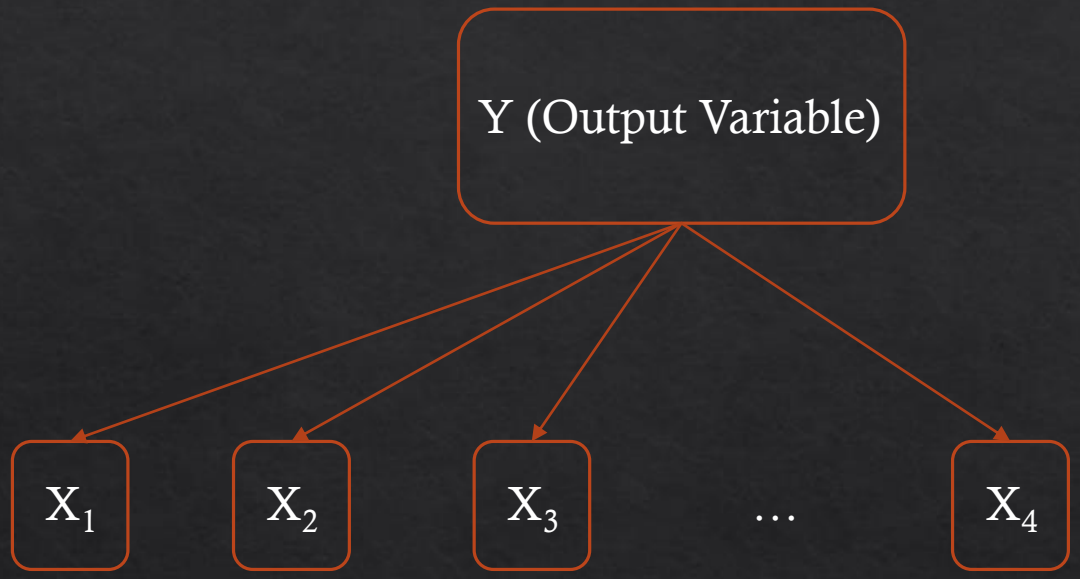
- i.e., normal distribution

To estimate the mean of $P(X_i \mid Y = y_k)$ we use:

$$\mu'_{ik} = \frac{\sum_j X_i^j * \delta(Yj = yk)}{\sum_j \delta(Yj = yk)}$$

Note that delta function is simply 1 when we have a training row with correct output variable and 0 otherwise

Continuous Features



So, each feature Xi is assumed to have a Gaussian Distribution

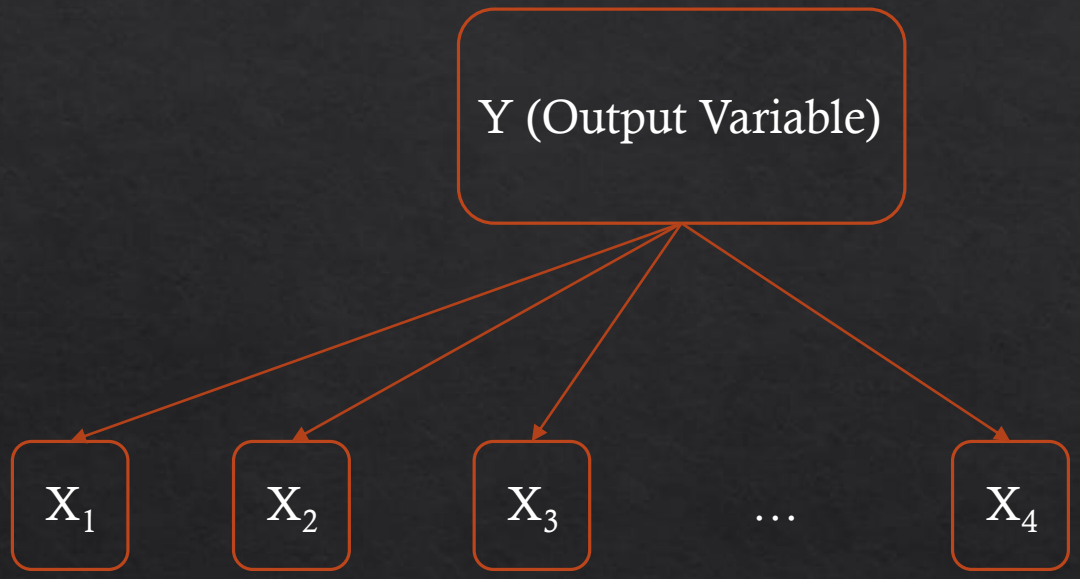
- i.e., normal distribution

To estimate the variance of $P(X_i \mid Y = y_k)$ we use:

$$\sigma^{2\prime}{}_{ik} = \frac{\sum_j (X^j{}_i - \mu'{}_{ik})^2 * \delta(Yj = yk)}{\sum_j \delta(Yj = yk)}$$

Note that delta function is simply 1 when we have a training row with correct output variable and 0 otherwise

Estimating P(Y)



Lastly, we need to estimate P(Y). We either do it like before:

$$\pi'_k = P'(Y = yk) = \frac{\#D\{Y = yk\}}{|D|}$$

So, #D{Y = y_k} is the number of rows where the output is exactly y_k

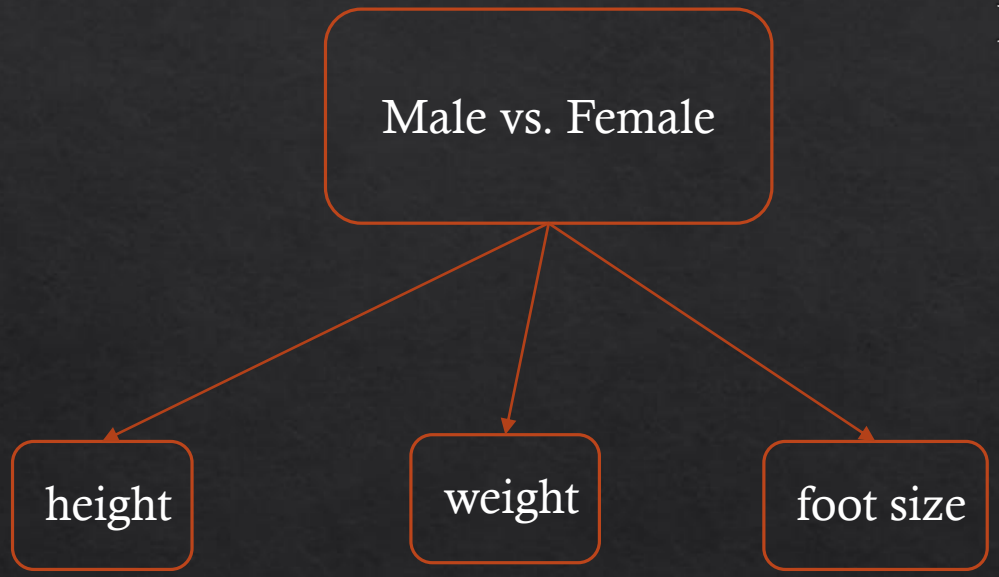
OR we use some prior knowledge about the underlying distribution

For Example: If output variable is gender

we might assume P(male) = 0.5

regardless of training set

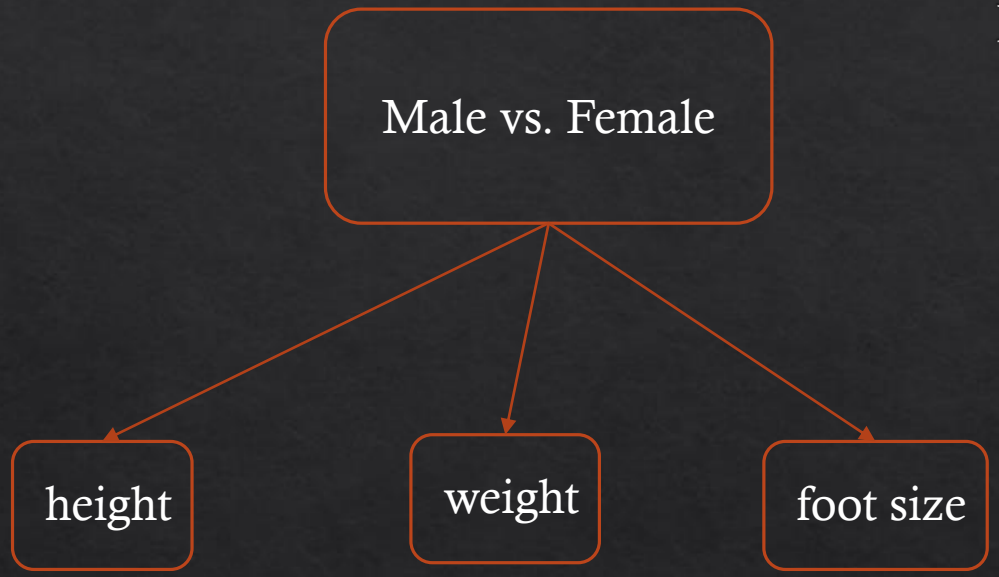
Example: Predicting Gender



Let's look at a specific example: From Wikipedia

sex	height (feet)	weight (lbs)	foot size(inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9

Example: Predicting Gender

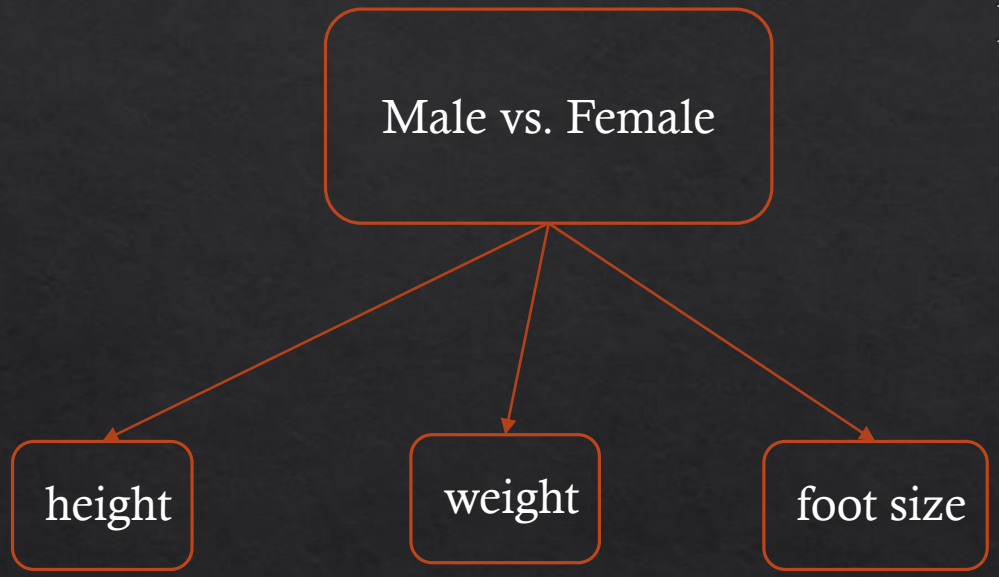


Let's look at a specific example: From Wikipedia

*These computed using mean and variance formulas from earlier

sex	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033e-02	176.25	1.2292e+02	11.25	9.1667e-01
female	5.4175	9.7225e-02	132.5	5.5833e+02	7.5	1.6667e+00

Example: Predicting Gender



Let's look at a specific example: From Wikipedia

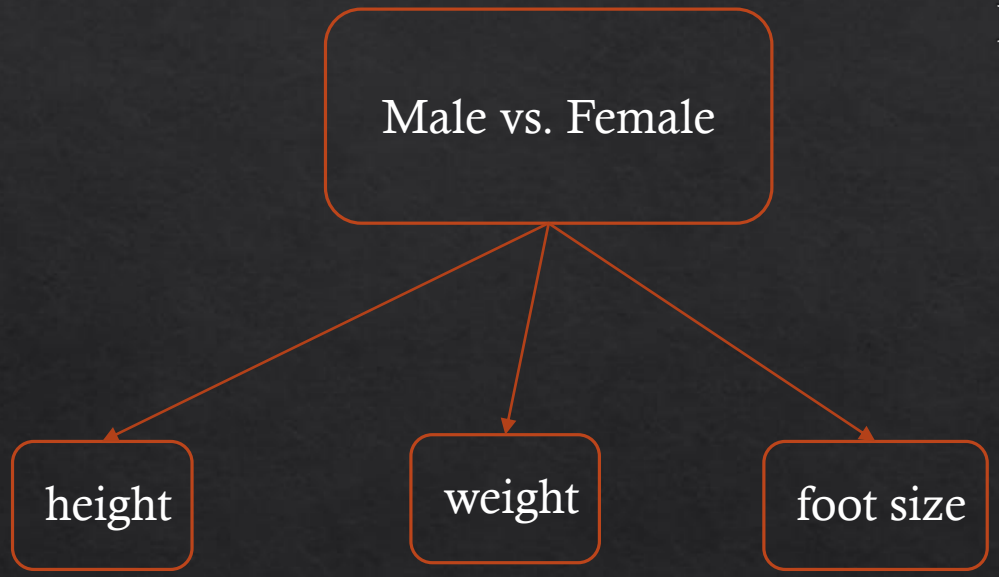
sex	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

*Now we want to estimate the gender of this new sample

*These computed using mean and variance formulas from earlier

sex	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033e-02	176.25	1.2292e+02	11.25	9.1667e-01
female	5.4175	9.7225e-02	132.5	5.5833e+02	7.5	1.6667e+00

Example: Predicting Gender



Let's look at a specific example: From Wikipedia

sex	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

*Now we want to estimate the gender of this new sample

$$posterior(male) = \frac{P(male)\, p(height|male)\, p(weight|male)\, p(footsize|male)}{evidence}$$

$$posterior(female) = \frac{P(female)\, p(height|female)\, p(weight|female)\, p(footsize|female)}{evidence}$$

*Compute these and see which is larger. Formula from graph structure

Example: Predicting Gender

sex	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

*Now we want to estimate the gender of this new sample

Need to be able to compute, for example:

P(Height | Male) from mean and variance

Probability density given by:

$$P(V \mid C) = \frac{\exp\left(-\frac{(V-\mu_c)^2}{2\sigma^2{}_c}\right)}{\sqrt{2\pi\sigma^2{}_c}}$$

V here is the value of the evidence in our new test data example (from the table to the left)

Example: Predicting Gender

sex	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

*Now we want to estimate the gender of this new sample

$$p(\text{height}|\text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(6-\mu)^2}{2\sigma^2}\right) \approx 1.5789$$

The 6 in equation here comes from test data under height

The mean and variance comes from mean and var. of height given in the table below (computed earlier)

Need to be able to compute, for example:

P(Height | Male) from mean and variance

Probability density given by:

$$P(V \mid C) = \frac{\exp(-\frac{(V-\mu_c)^2}{2\sigma^2{}_c})}{\sqrt{2\pi\sigma^2{}_c}}$$

sex	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033e-02	176.25	1.2292e+02	11.25	9.1667e-01
female	5.4175	9.7225e-02	132.5	5.5833e+02	7.5	1.6667e+00

Example: Predicting Gender

sex	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

*Now we want to estimate the gender of this new sample

Example: Predicting Gender

sex	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

*Now we want to estimate the gender of this new sample

Given these estimated mean and variance values

sex	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033e-02	176.25	1.2292e+02	11.25	9.1667e-01
female	5.4175	9.7225e-02	132.5	5.5833e+02	7.5	1.6667e+00

Example: Predicting Gender

sex	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

*Now we want to estimate the gender of this new sample

Given these estimated mean and variance values

sex	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033e-02	176.25	1.2292e+02	11.25	9.1667e-01
female	5.4175	9.7225e-02	132.5	5.5833e+02	7.5	1.6667e+00

$$posterior(male) = \frac{P(male)\, p(height|male)\, p(weight|male)\, p(footsize|male)}{evidence}$$

$$posterior(female) = \frac{P(female)\, p(height|female)\, p(weight|female)\, p(footsize|female)}{evidence}$$

*Compute these and see which is larger. Formula from graph structure

Notice that evidence is same for both, so no need to compute

sex	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

Example: Predicting Gender

$$posterior(male) = \frac{P(male)\, p(height|male)\, p(weight|male)\, p(footsize|male)}{evidence}$$

$$posterior(female) = \frac{P(female)\, p(height|female)\, p(weight|female)\, p(footsize|female)}{evidence}$$

$p(\text{weight}|\text{male}) = 5.9881 \cdot 10^{-6}$
$p(\text{foot size}|\text{male}) = 1.3112 \cdot 10^{-3}$
$\text{posterior numerator (male)} = \text{their product} = 6.1984 \cdot 10^{-9}$
$P(\text{female}) = 0.5$
$p(\text{height}|\text{female}) = 2.2346 \cdot 10^{-1}$
$p(\text{weight}|\text{female}) = 1.6789 \cdot 10^{-2}$
$p(\text{foot size}|\text{female}) = 2.8669 \cdot 10^{-1}$
$\text{posterior numerator (female)} = \text{their product} = 5.3778 \cdot 10^{-4}$

Posterior value for male is lower than for female, so we predict this is a female.