Introduction to Machine Learning

Lecture 4

Instructor:
Dr. Tom Arodz

ML: Typical assumptions

- The modeled phenomenon is poorly understood / too complex to simulate
- The features are somewhat informative but not perfectly correlated with the class
- The association between *regions of feature space* and the *class variable* is fixed
- The association between features and class we can learn is likely to be accurate only for objects similar to our training set

- These assumptions lead to a probabilistic view of ML

ML and probability

- Example

Feature X1 – height (0-very short, 1-short, 2-medium)

Class Y – tribe (0-hobbit, 1-dwarf)

ML and probability

Feature X1 – height (0-v. short, 1-short, 2-medium)

Class Y – tribe (0-hobbit, 1-dwarf)

- What we study is rarely deterministic / crisp
  - Not true that: *all hobbits are short or v.short, all dwarves are medium*

- Probability comes into play in several ways:
  - Features and dataset:
    - v. short folks are 30% of the population of Middle-earth
    - In our specific training set, v. short folks are 27%, not 30%
  - Class (y) vs features (x):
    - y=>x: if *class is hobbit*,
      then height is v.short 40% of the time,
      short 35% of the time
    - x=>y: if *height is v. short*,
      then it is a hobbit 85% of the time

ML and probability

Feature X1 – height (0-v. short, 1-short, 2-medium)

Class Y – tribe (0-hobbit, 1-dwarf)

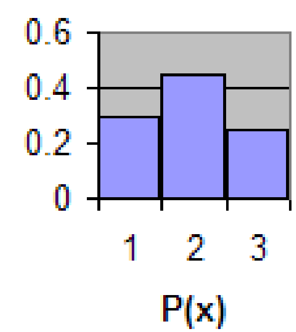
- Probability comes into play in several ways:
  - Features and dataset:
    - v. short folks are 30% of the population
      - Probability(X1 = 0)=0.3, P(X1=1)=0.45, …
      - **Probability distribution** D for X1:
        - 0:0.3, 1:0.45, 2:0.25
    - In our specific training set, v. short folks are 27%, not 30%
      - Dataset comes to us by randomly drawing from **distribution** D over features
  - Class vs features:
    - y=>x if hobbit, then height is v.short 40% of time
      - Probability of X1=0 given Y=0:
conditional probability: $P(X1=0 \mid Y=0) = 0.4$
    - x=>y if height is v. short, then it is a hobbit 85% of time
      - Probability of Y=0 given X1=0:
conditional probability: $P(Y=0 \mid X1=0) = 0.85$

ML and probability

- *Sample space*: spectrum of possible observations
  - E.g. sample space for dice = {1,2,3,4,5,6}
  - E.g. feature X1 can be: 0,1,2
  - E.g. class Y can b: 0, 1
- *Event space*: sets of observations
  - E.g. "*1 on a dice*", "*even number on a dice"*
  - E.g. observed feature value: 0,
    observed feature value: 1 or 2
- *Probability* : function that assigns a number in [0,1] range to events:
  - P(*event*) quantifies the degree of our belief that *event* happens (e.g. equality is true)
  - E.g. P(feature=0), P(feature=1 or feature=2),P(class=1)

ML and probability

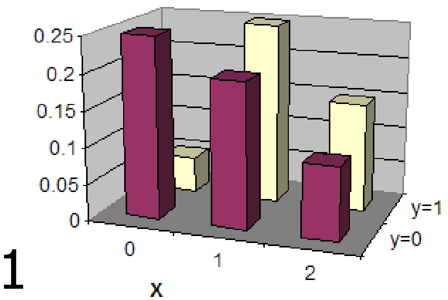
- *Sample space*: possible observations
  - E.g. feature X1 can be: 0,1,2
- We will see two types of sample spaces
  - Real numbers (typically individual features are reals)
    - We will call these *random variables*: e.g. X1, Y
    - We define *probability distribution over random variable*, e.g. P(X1=0)=0.25, P(Y=1)=0.4
    - When we talk about distribution as a whole (not probability for specific value, P(X1=0)) in ML we often use D to denote the distribution
      - We write *X1 ~ D* to say values of X1 are distributed according to distrib. *D*
        - P(X1=0) = *D*(0), P(X1=1)=*D*(1), …
    - Distribution D(x) gives us probability values P(X1=x) for each possible observation x
      - These values are in [0,1] range, and add up to 1



ML and probability

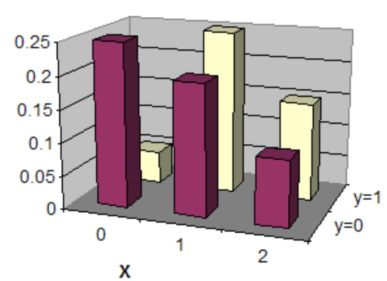
- *Sample space*: possible observations

- We will see two types of sample spaces
  - Real numbers (e.g. individual features)
  - Vectors (*multiple features*, and/or *features + class*)
    - *Joint distribution over multiple random variables*
      - Over all possible combinations of values
    - E.g. P(X1=0, Y=1)=0.2
      - the probability that value of X1 will be equal to 0 AND value of Y will be equal to 1
      - Again, we use D to denote distribution itself
        - (x,y) ~ D
        - P(X1=0, Y=1)=D(0,1)
      - Again, values of D are in [0,1] range and add up to 1 over the whole set of distinct possibilities



Notation

- P(X=x) is the probability that variable X assumes value x
- Often, we use simplified notation, with variable implied by context:
  - E.g. P(+1) instead of P(Y=+1) if it's clear we are talking about class
  - E.g. P(y) or P(Y) to talk about the probability of classes in general, not of specific class values like +1
- Distributions
  - We often use subscript to denote which distribution we mean
    - $X \sim D_x$ $Y \sim D_y$ $Y \sim D_{y|x}$
- We often use z=(x,y) to denote all features and class, jointly
  - D_z is the distribution over those, it gives us $P(X1=0, Y=1)=D_z(0,1)$
    - $z \sim D_z$,
    - $(x,y) \sim D_z$



ML and probability

- Back to Middle-earth

Feature X1 – height (0-very short, 1-short, 2-medium)

Class Y – tribe (0-hobbit, 1-dwarf)

We have two separate distributions,
over the feature(s)
$X1 \sim D_{x1}$

and over the class

$Y \sim D_y$

P(x1=0)	0.3
P(x1=1)	0.45
P(x1=2)	0.25

P(y=0)	0.55
P(y=1)	0.45

The distribution over the feature(s) covers possible values of features from which our samples come from

The distribution over the classes is typically a discrete distribution over just two possibilities (+1/-1 or 1/0):
$P(+1) + P(-1) = 1$ or $P(+1) + P(0) = 1$

Using probability for predicting

- If we know the probability distribution for individual random variables (features P(x) and class P(Y)), does it help making class predictions?

P(x1=0)	0.3
P(x1=1)	0.45
P(x1=2)	0.25

P(y=0)	0.55
P(y=1)	0.45

Feature X1 – height (0-very short, 1-short, 2-medium)

Class Y – tribe (0-hobbit, 1-dwarf)

If you know the above distrib's,
and that x1=short
what would you predict?

if you knew x1=v.short,
What would you predict?

ML and probability

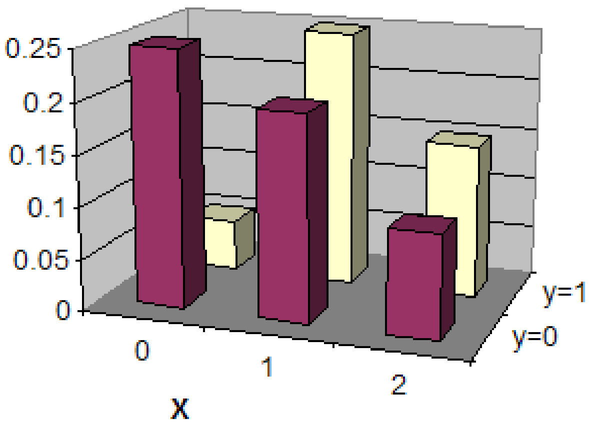
- Back to Middle-earth

Classes probabilities are not independent from feature probs.
(we learn to use that relationship to make predictions).

P(x1=0)	0.3
P(x1=1)	0.45
P(x1=2)	0.25

P(y=0)	0.55
P(y=1)	0.45

Joint distribution D over z=(X1,Y):



	D
P(x1=0,y=0)	0.25
P(x1=0,y=1)	0.05
P(x1=1,y=0)	0.2
P(x1=1,y=1)	0.25
P(x1=2,y=0)	0.1
P(x1=2,y=1)	0.15

How can joint probability distribution over (features, class) vectors help in making predictions?

ML and probability

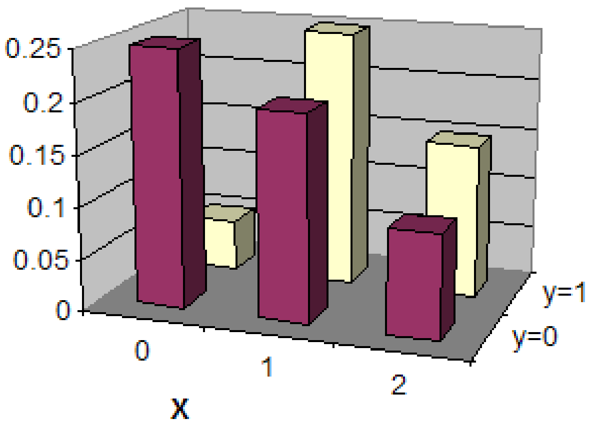
P(x1=0)	0.3
P(x1=1)	0.45
P(x1=2)	0.25

P(y=0)	0.55
P(y=1)	0.45

- Back to Middle-earth

  Classes probabilities are not independent from feature probs.
  (we learn to use that relationship to make predictions).

  Joint distribution D over z=(X1,Y):



	D
P(x1=0,y=0)	0.25
P(x1=0,y=1)	0.05
P(x1=1,y=0)	0.2
P(x1=1,y=1)	0.25
P(x1=2,y=0)	0.1
P(x1=2,y=1)	0.15

How can joint probability distribution over (features, class) vectors help in making predictions?

If we know (x1=0), we can look up whether P(x1=0,y=0) or P(x1=0,y=1) is higher

Basically, make prediction using *conditional probability*!

Using probability for predicting

- What can help us directly in making predictions is *conditional probability* of class given features

  - *Conditional probability of Y given X*: P(Y=y | X=x) is the probability that Y will be equal to y if we know that X took the value of x
    - Often, we just write P(y|x)

  - We see a v.short character (X1=0), is it a hobbit (Y=0) or a dwarf (Y=1)?
    - Probability that class is 0 if we know height is "v. short" P(Y=0 | X1=0) = 0.83
    - Probability that class is 1 if we know height is "v. short" P(Y=1 | X1=0) = 0.17
    - What should our prediction be? Hobbit (0)!

What is *conditional probability*?

- What can help us directly in making predictions is *conditional probability* of class given features
  - *Conditional probability of Y given X*: P(Y=y | X=x) is the probability that Y will be equal to y if we know that X took the value of x

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

  - We can derive conditional probability from joint probability

    P(Y=y | X=x) = P(Y=y , X=x) / P(X=x)

  - E.g. *X=1* happens 50 of 100 times, and *Y=1,X=1* happens 20 out of 100 times.
    - When X=1 happens (50 times)
    - Y=1 happens 20 times out of the 50 times
    - P(Y=1|X=1)=20/50 =0.4

Probability so far - recap

- P(A) – probability of event A happening
  - P(temp<32F)=0.1
- P(AB) – joint probability of both A and B events happening
  - P(temp<32F, snow)=0.05
- P(A|B) – probability of A happening if B is happening
  - P(snow | temp<32F) = 0.5
  - P(temp<32F | snow) = 0.95
- P(A|B) = P(AB) / P(B)
  - P(snow | temp<32F)=P(temp<32F, snow) / P(temp<32F)
    0.5 = 0.05 / 0.1
- P(A|B)*P(B)=P(AB)
  - P(snow | temp<32F)*P(temp<32F)=P(temp<32F, snow)
    0.5*0.1 = 0.05
- We don't know P(snow), can we deduce it?

Probability so far - recap

- P(A) – probability of event A happening
  - P(temp<32F)=0.1
- P(AB) – joint probability of both A and B events happening
  - P(temp<32F, snow)=0.05
- P(A|B) – probability of A happening if B is happening
  - P(snow | temp<32F) = 0.5
  - P(temp<32F | snow) = 0.95
- P(A|B) = P(AB) / P(B)
  - P(snow | temp<32F)=P(temp<32F, snow) / P(temp<32F)
    0.5 = 0.05 / 0.1
- P(A|B)*P(B)=P(AB)
  - P(snow | temp<32F)*P(temp<32F)=P(temp<32F, snow)
    0.5*0.1 = 0.05
- We don't know P(snow), can we deduce it?
  - P(temp<32F | snow) * P(snow) = P(temp<32F,snow)
    0.95*??? = 0.05
    we can deduce that P(snow)=0.0526

Using probability for predicting

- Predictions from conditional probability P(y|x)

Feature X1 –height (0-v. short, 1-short, 2-medium)

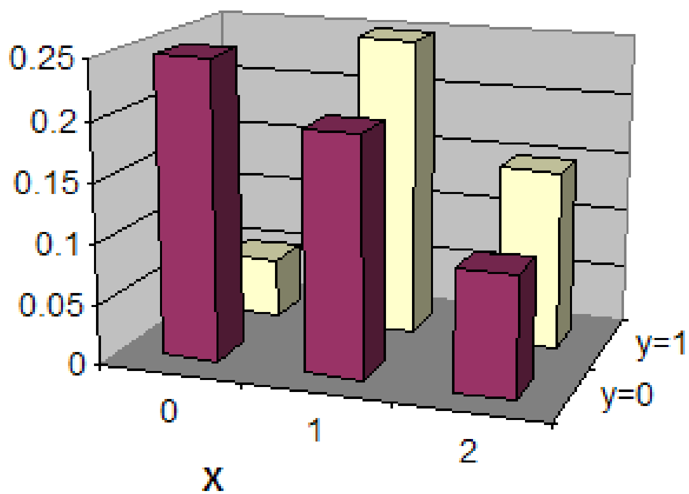
Class Y – tribe (0-hobbit, 1-dwarf)

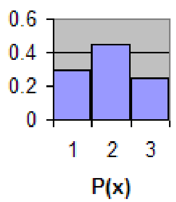
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

	D
P(x1=0,y=0)	0.25
P(x1=0,y=1)	0.05
P(x1=1,y=0)	0.2
P(x1=1,y=1)	0.25
P(x1=2,y=0)	0.1
P(x1=2,y=1)	0.15

P(y=0)	0.55
P(y=1)	0.45

P(x1=0)	0.3
P(x1=1)	0.45
P(x1=2)	0.25





$D_{Y\|X}$		
P(y=0|x1=0)	0.833333	0.25/0.3
P(y=1|x1=0)	0.166667	0.05/0.3
P(y=0|x1=1)	0.444444	0.2/0.45
P(y=1|x1=1)	0.555556	0.25/0.45
P(y=0|x1=2)	0.4	0.1/0.25
P(y=1|x1=2)	0.6	0.15/0.25

For each value of x (features), predict the most probable value of y (class)

Using probability for predicting

- Predictions from conditional probability P(y|x)

Feature X1 –height (0-v. short, 1-short, 2-medium)

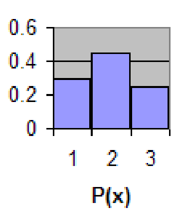
Class Y – tribe (0-hobbit, 1-dwarf)

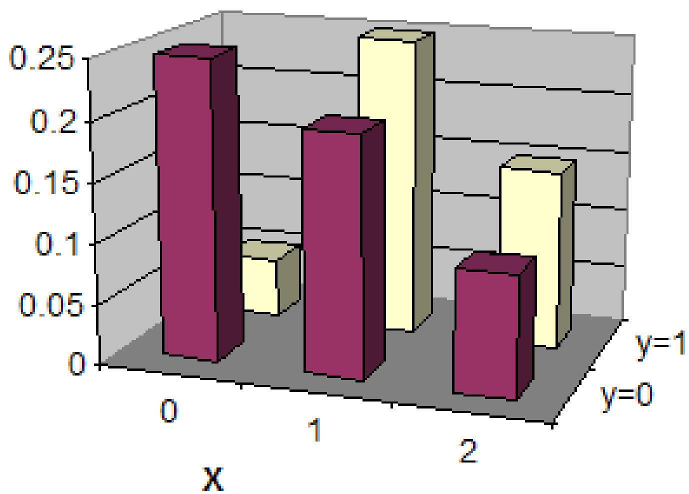
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

	D
P(x1=0,y=0)	0.25
P(x1=0,y=1)	0.05
P(x1=1,y=0)	0.2
P(x1=1,y=1)	0.25
P(x1=2,y=0)	0.1
P(x1=2,y=1)	0.15

P(y=0)	0.55
P(y=1)	0.45

P(x1=0)	0.3
P(x1=1)	0.45
P(x1=2)	0.25





$D_{Y\|X}$		
P(y=0|x1=0)	0.833333	0.25/0.3
P(y=1|x1=0)	0.166667	0.05/0.3
P(y=0|x1=1)	0.444444	0.2/0.45
P(y=1|x1=1)	0.555556	0.25/0.45
P(y=0|x1=2)	0.4	0.1/0.25
P(y=1|x1=2)	0.6	0.15/0.25

Is it possible to avoid incorrect predictions?

What is the probability of making a wrong prediction?

Using probability for predicting

- Predictions from conditional probability P(y|x)

Feature X1 –height (0-v. short, 1-short, 2-medium)

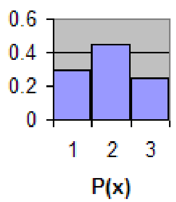
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Class Y – tribe (0-hobbit, 1-dwarf)

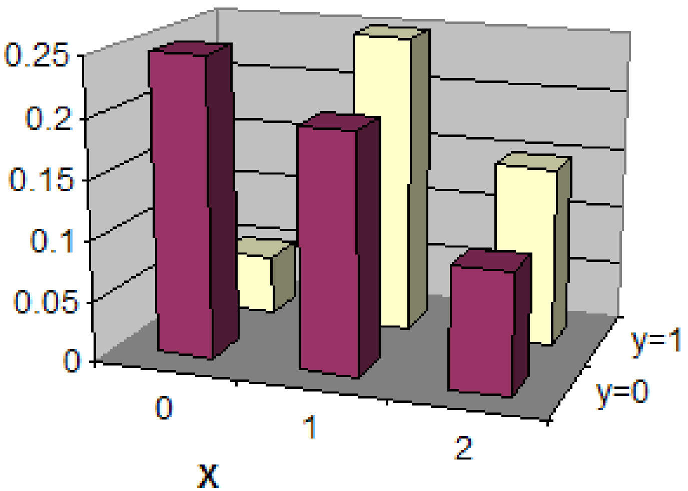
	D
P(x1=0,y=0)	0.25
P(x1=0,y=1)	0.05
P(x1=1,y=0)	0.2
P(x1=1,y=1)	0.25
P(x1=2,y=0)	0.1
P(x1=2,y=1)	0.15

P(y=0)	0.55
P(y=1)	0.45

P(x1=0)	0.3
P(x1=1)	0.45
P(x1=2)	0.25

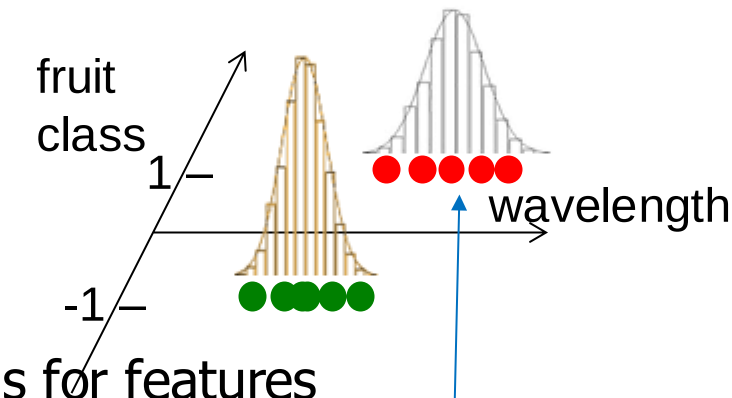


Expected error is:
0.05+0.2+0.1=0.35=35%



Continuous features

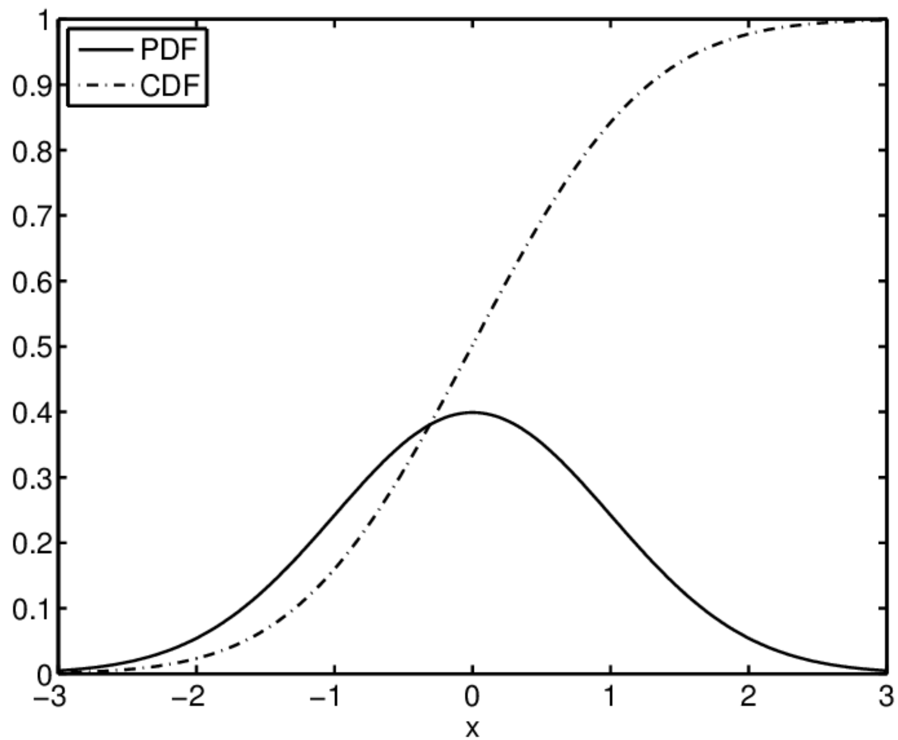
- Apple vs Orange
  - **Joint distribution over (*wavelength* x *fruit class*)**



- Here, class is still discrete (-1,+1)
  But feature is continuous (real number)
- Our distributions will typically be over reals for features
  - Most often, probability of any specific real value is 0
    - E.g., probability of height (not rounded) being exactly 5.678901234 ft is 0
  - But, probability of a range of values is typically >0
    - E.g. probability of height in [5.67-5.68] ft
  - For continuous variables, we have *probability density function* (pdf)
    - Intuitively, pdf p(x) is a function that tells us
      the probability of seeing values from a very small region around x
      relative to other regions
      - As if we did histograms with narrower and narrower bins,
        always using infinite number of samples

Continuous features (grad only)

- For continuous variables, we have *probability density function* (pdf)
  - Intuitively, pdf p(x) is a function that tells us the probability of seeing values from a very small region around x relative to other regions
- We first define *cumulative distribution function* CDF(x)
  - CDF(x) = P(X<=x)
- Then, define PDF from CDF
  - PDF is the derivative of CDF with respect to x
    pdf(x) = *d* CDF / *d*x



Probabilistic setting - summary

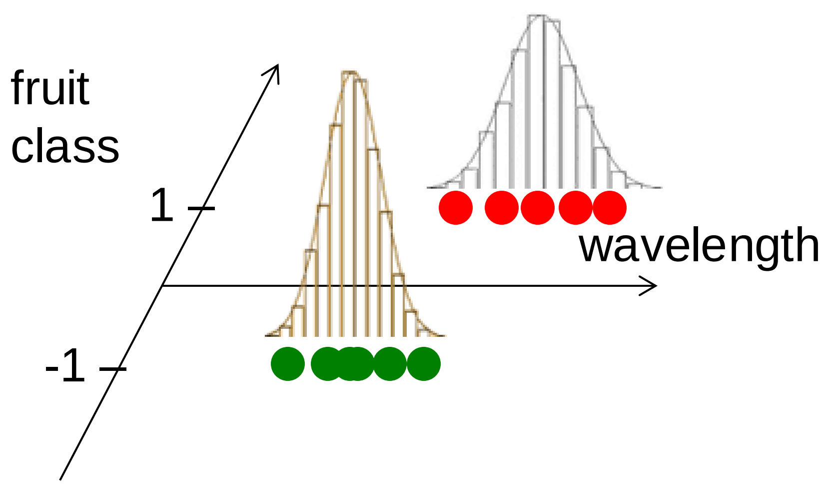
- We have observations in a fixed F-dimensional feature space X
  - Every sample x is a vector (point) in that feature space

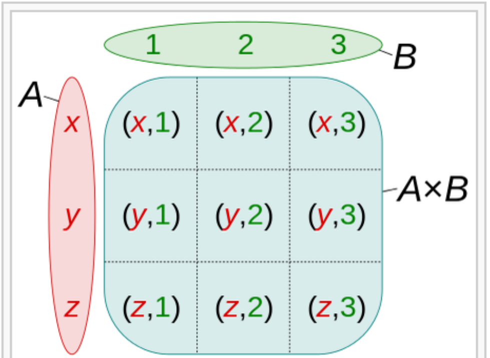
$$\mathbf{x} = \left[x^{\langle 1 \rangle}, x^{\langle 2 \rangle}, \ldots, x^{\langle F \rangle}\right]^T \quad \mathbf{x} \in \mathcal{X} \quad \mathcal{X} \subset \mathbb{R}^F$$

- Sample x belongs to class y, {-1, +1} (or {0,1}, or {1,2,3,..})
  - So together we have an extended space

$$\mathbf{z} = (\mathbf{x}, y) \qquad \mathcal{Z} = \mathcal{X} \times \{-1, +1\}$$

Cartesian product



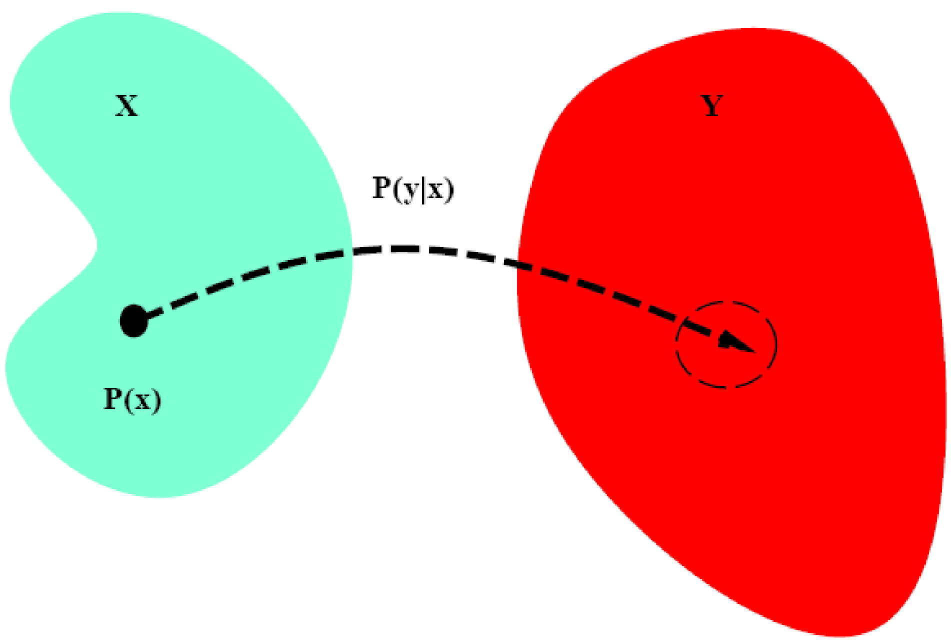


Cartesian product $A \times B$ of the sets $A = \{x, y, z\}$ and $B = \{1, 2, 3\}$

Probabilistic setting - summary

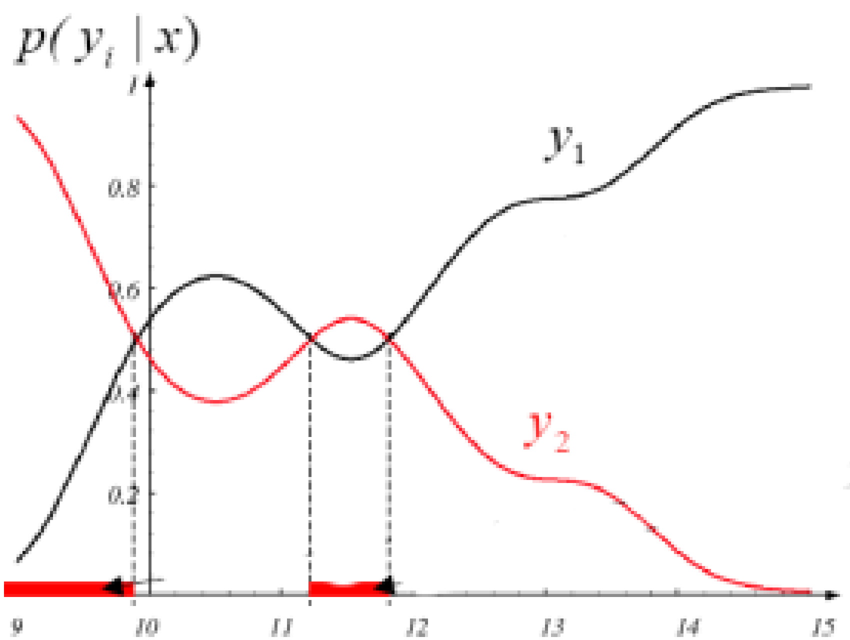
- Examples come from space $\mathcal{Z} = \mathcal{X} \times \{-1, +1\}$ $\mathbf{z} = (\mathbf{x}, y)$
- Over that space, we have a joint probability distribution
- Samples are obtained from that distribution and have probability P(z) = P(x,y)
- We can factor it using conditional probability to separate P(x) from P(y|x)
  - P(z)= P(x,y)=P(y|x)P(x)

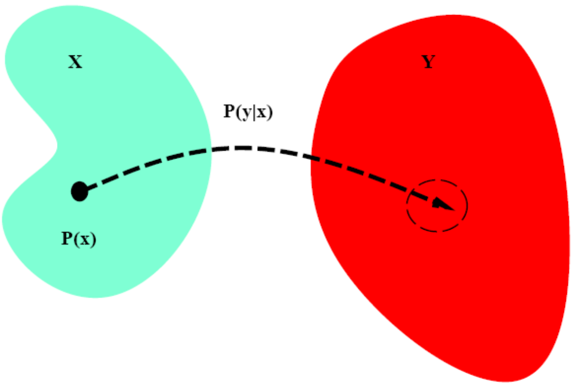
P(y|x) = conditional probability
prob. of seeing class y
if we're observing sample x



Probabilistic setting - summary

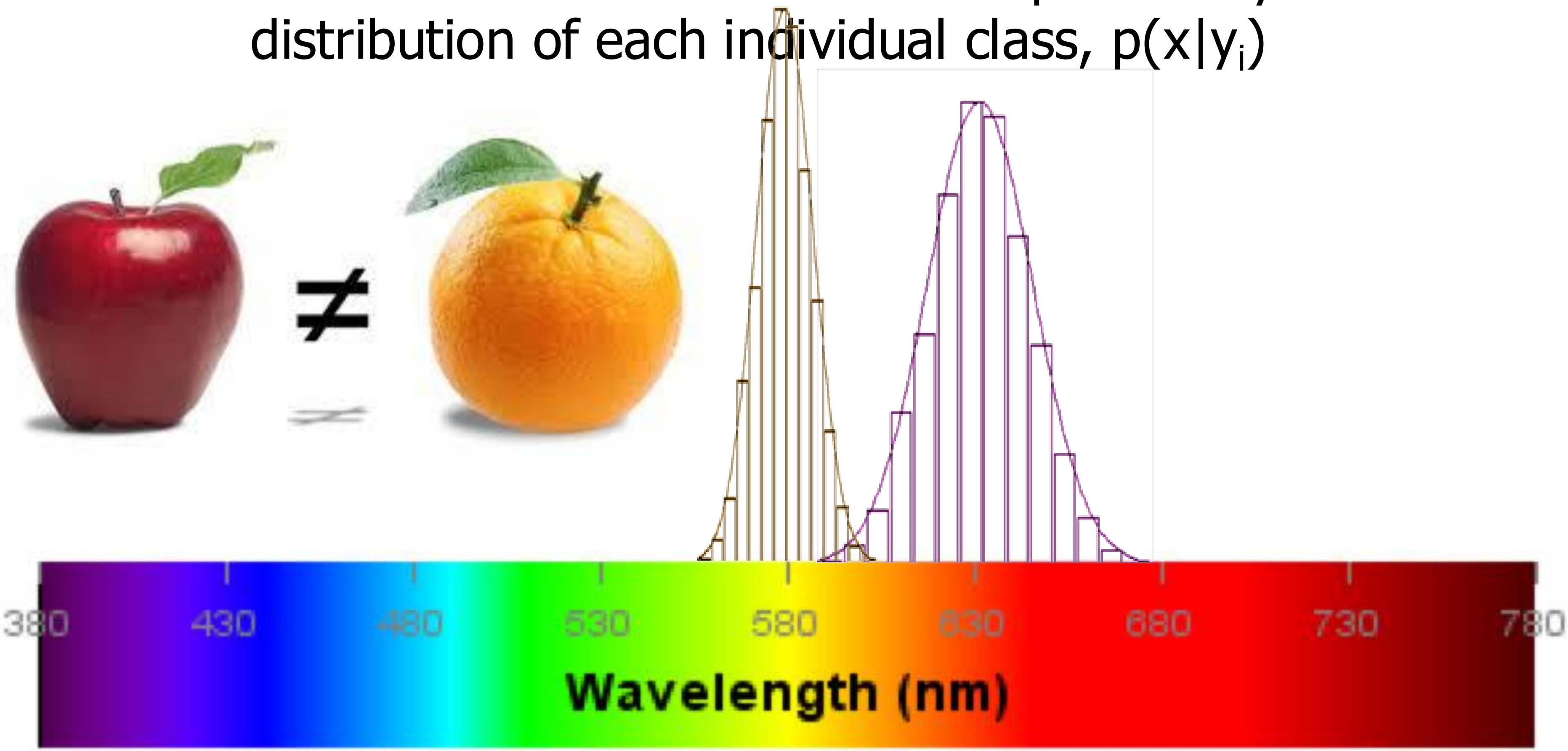
- What we want for classification is p(y|x):
what is the most probably class y for a given x?
- For each value of x (features),
predict the most probable
value of y (class)





Probabilistic setting - summary

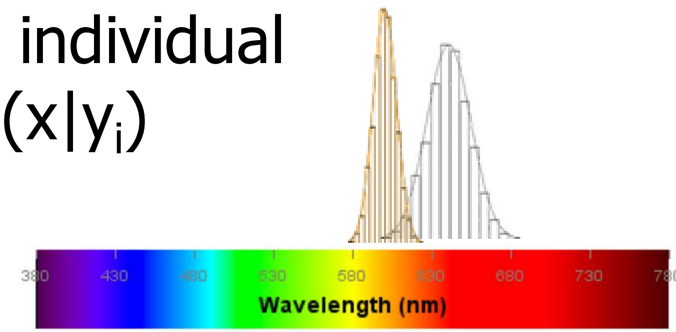
- It is often much easier to obtain probability distribution of each individual class, $p(x|y_i)$

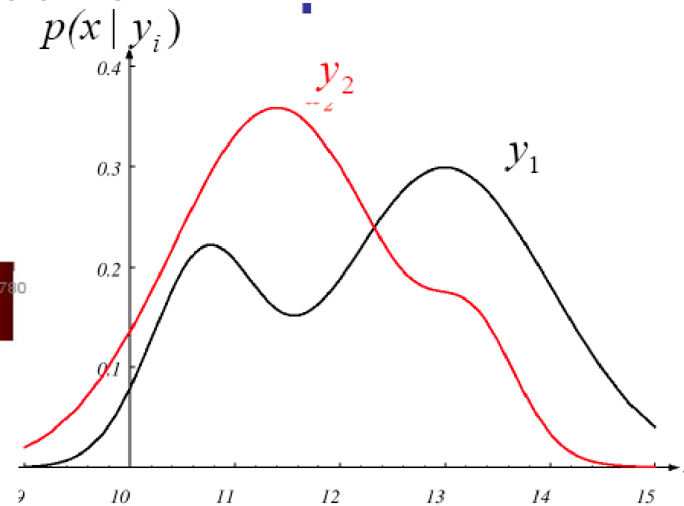


- **Probability distribution of wavelength for Apples**
- **Probability distribution of wavelength for Oranges**

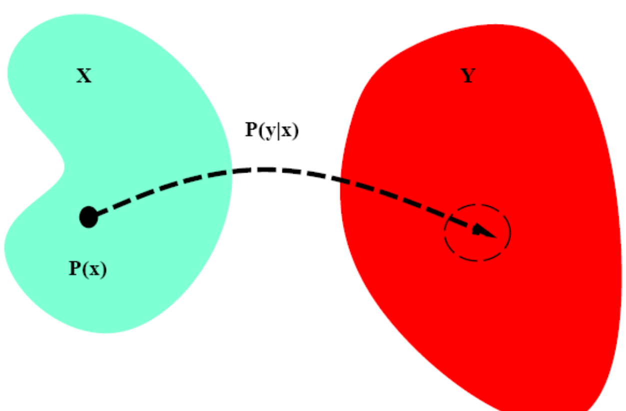
Probabilistic setting - summary

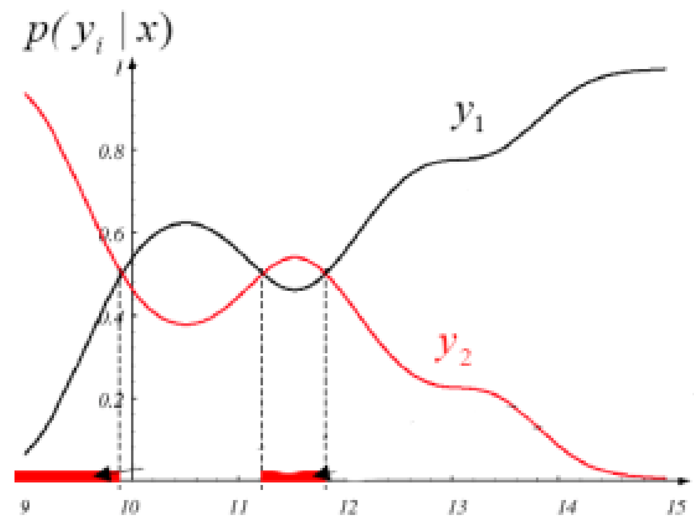
- We may have probability distribution of each individual class, $p(x|y_i)$



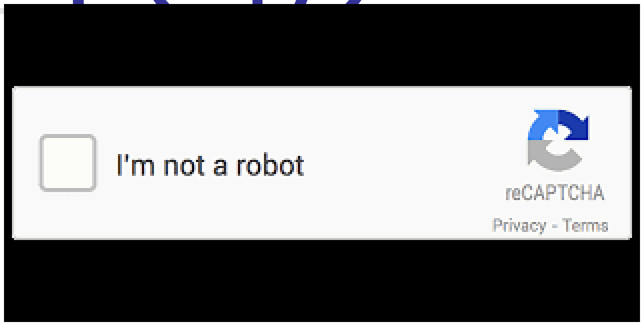


- But what we really want for classification is $p(y_i|x)$: *what is the probability of class y_i for given value of x?*





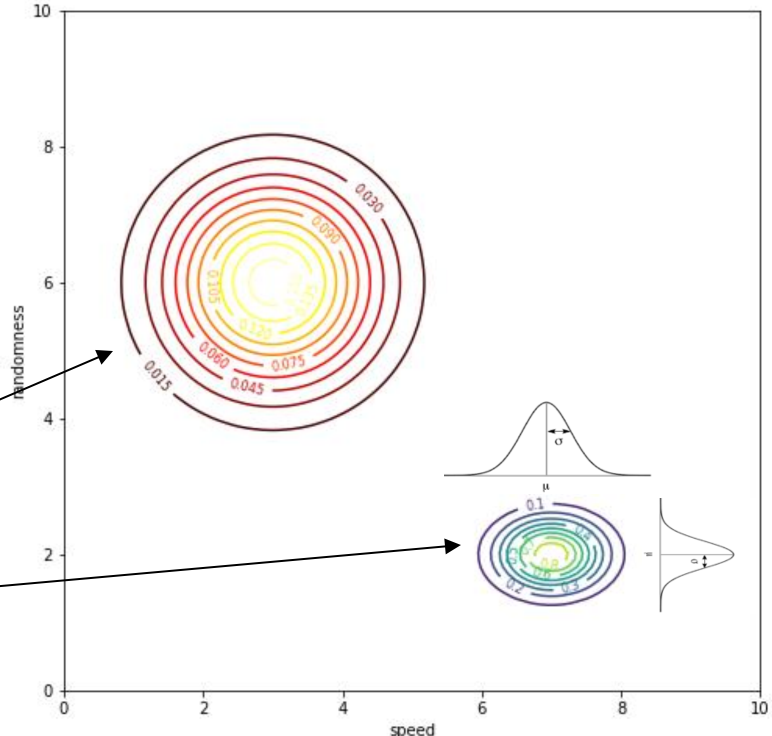
Often, we just have p(x|y)



Two classes: robot or human

Two features, $x=(x^1,x^2)$:

- speed of moving mouse cursor to click
- randomness (deviation from straight line) of the move



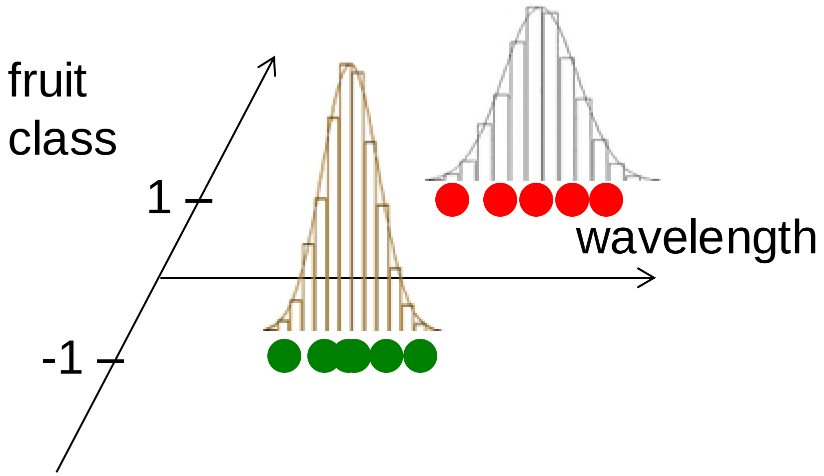
p(x | human)

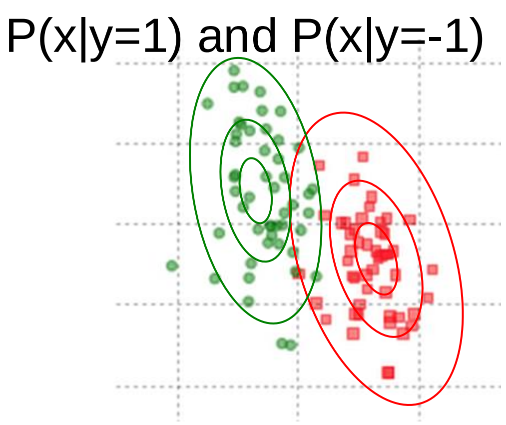
p(x | robot)

Probabilistic decision making

- Assume:
  - Somehow we got to know:
    - The distributions $P(x|y_i)$ for each class y_i (i.e., the distributions over feature vectors x)
    - The probabilities $P(y_i)$ for each class y_i (i.e., single numbers)

  - How do we make decisions given this information?





Bayes theorem

- Conditional probability Y given X
  $P(Y=y \mid X=x) = P(X=x \text{ AND } Y=y) / P(X=x)$
  - That means: $P(X=x \text{ AND } Y=y) = P(Y=y \mid X=x)\, P(X=x)$
- Conditional probability X given Y
  $P(X=x \mid Y=y) = P(X=x \text{ AND } Y=y) / P(Y=y)$
  - That means: $P(X=x \text{ AND } Y=y) = P(X=x \mid Y=y)\, P(Y=y)$
- Bayes Theorem links these two conditional probabilities: $P(y|x) = P(x|y)\, P(y) / P(x)$
  - From $P(y|x)P(x)=P(x,y)=P(x|y)P(y)$

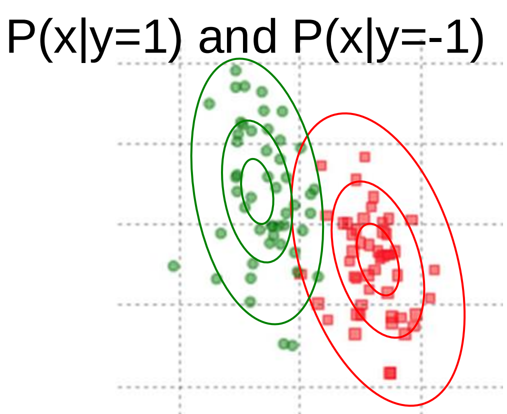
Probabilistic classification

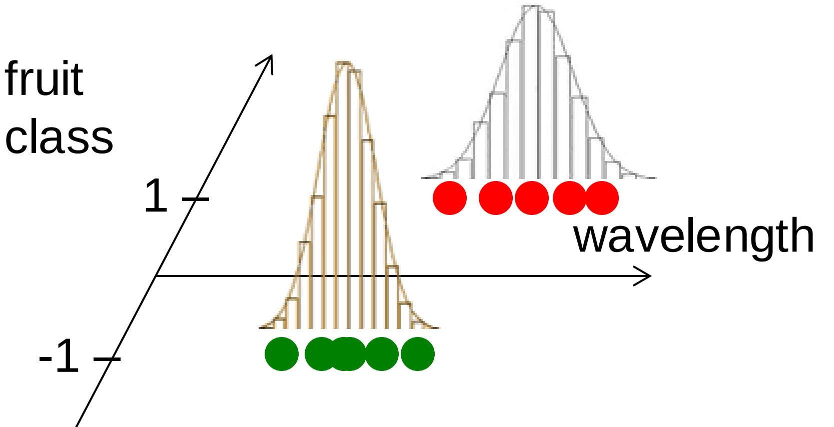
- Assume:
  - Somehow we got to know:
    - The distributions $P(x|y_i)$ for each class y_i (i.e., the distributions over feature vectors x)
    - The probabilities $P(y_i)$ for each class y_i (i.e., single numbers)
  - How do we make decisions given this information?

We use Bayes Theorem!

$$P(y_i \mid x) = P(x \mid y_i)P(y_i) \,/\, P(x)$$

$$= P(x \mid y_i)P(y_i) \,/\, \Sigma_i P(x \mid y_i)P(y_i)$$





Probabilistic classification

- Detailed derivation:

$P(A)=\Sigma_i P(A|B_i)P(B_i)$

- $\mathbf{p(y_i \mid x) = p(x \mid y_i)\, p(y_i) \,/\, p(x)}$
  - $P(y_i,x)=P(y_i|x)P(x)=P(y_i|x)P(x)$

    $P(y_i,x)=P(x|y_i)P(y_i)=P(x|y_i)P(y_i)$
- $\mathbf{p(y_i \mid x) = p(x \mid y_i)\, p(y_i) \,/\, \Sigma_i\, p(x \mid y_i)\, P(y_i)}$
  - $\Sigma_i p(x|y_i)P(y_i) = \Sigma_i p(x|y_i)P(y_i)$

    $=\Sigma_i p(x|y_i)P(y_i) = \Sigma_i p(x,y_i) = P(x)$

Recap: Probabilistic classification

- How do we make decisions given $P(x|y_i)$ and $P(y_i)$?
  - $p(y_i \mid x) = p(x \mid y_i)\, p(y_i) \,/\, p(x)$

    $\sim p(x \mid y_i)\, p(y_i)$

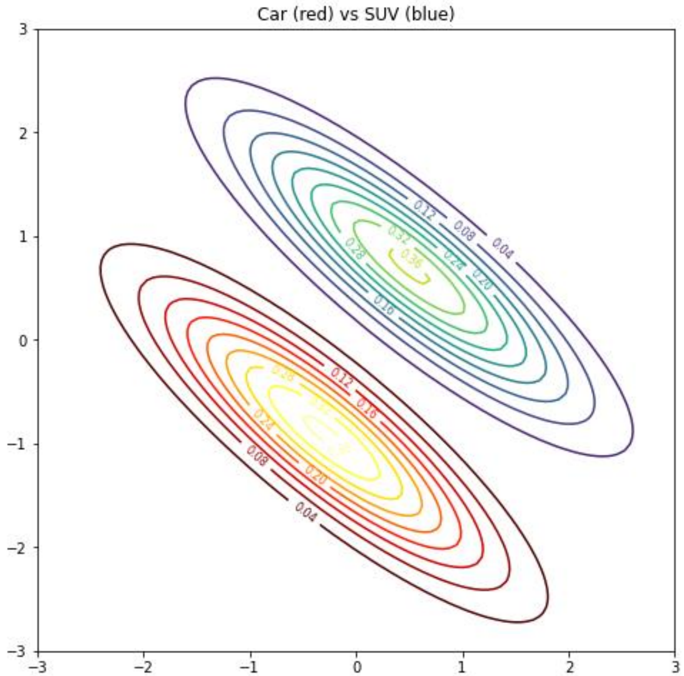
p(x) same for each y_i

- In python (using scipy + numpy):

```
from scipy.stats import multivariate_normal;
def predict_y0or1_for_x(x):
     # someone gave us means, covariances, and prob. of classes
     distr_x_given_class0 = multivariate_normal(mean=mean0, cov=covariance0)
     distr_x_given_class1 = multivariate_normal(mean=mean1, cov=covariance1)
     p_y_0 = 0.45;            p_y_1 = 1-py_0;
     # get p(x|y)
     p_x_given_y0 = distr_x_given_class0.pdf(x)
     p_x_given_y1 = distr_x_given_class1.pdf(x)
     # calculate p(y|x) // ignoring p(x)
     p_y0_given_x = p_y_0 * p_x_given_y0
     p_y1_given_x = p_y_1 * p_x_given_y1
     if (p_y0_given_x > p_y1_given_x):          return 0;
     else:                                      return 1;
```

Probabilistic classification

Two classes: car or SUV (HW1 data)



- How do we make decisions given $P(x|y_i)$ and $P(y_i)$?
  - **p(car | x) = p(x | car) p(car)**
  - **p(SUV | x) = p(x | SUV) p(SUV)**

- In python:

```
distr_x_given_car = multivariate_normal
                        (mean=meanCar, cov=covarianceCar)
dist_x_given_SUV = multivariate_normal
                        (mean=meanSUV, cov=covarianceSUV)
```

- when new vehicle pops up, we just do the arithmetic to make a prediction