Lesson 7

Bayes Classifier

Awareness through analysis and synthesis

WHOLENESS OF THE LESSON

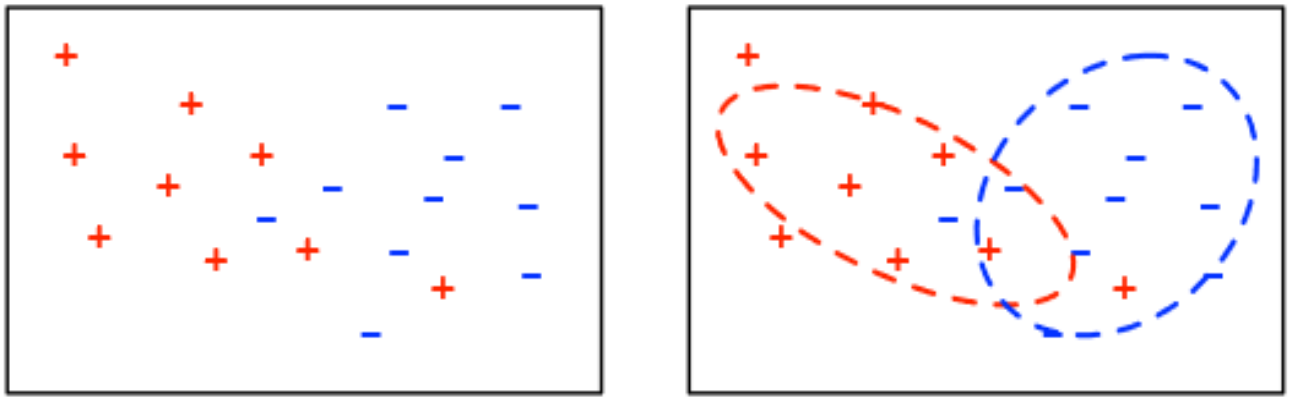
Bayes classifier is based on probability theory and hence gives a scientific foundation for the classifier created. The classifier always pick the item that has the highest probability by first modeling the problem using a joint distribution function.

Likewise, as discussed in SCI, more successful action results from a deeper dive into silence, into pure intelligence, just as, in archery, the arrow flies truer and hits its mark more consistently if it is pulled back farther on the bow.

Generative Approach

- Based on probability distributions.
- the main idea with the generative approach is to **fit each class separately with a probability distribution.**

Generative Approach



The learning process:

- Fit a probability distribution to each class, individually

To classify a new point:

- Which of these distributions was it most likely to have come from?

Example

- There are three classes labeled 1, 2, and 3.
- Data is just one feature: A real number.
- First consider all points whose label is 1.
  - fit a probability distribution function $P_1(x)$.
- Then consider all points whose label is 2.
  - fit a probability distribution function $P_2(x)$.
- Then consider all points whose label is 3.
  - fit a probability distribution function $P_3(x)$.

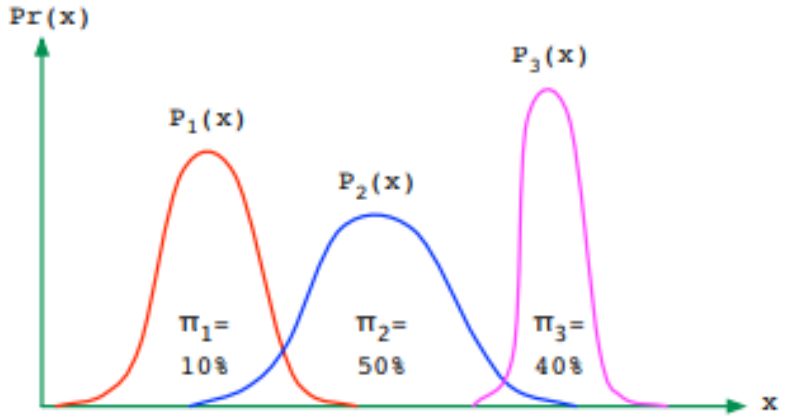
Example

- Assume label 1 appear 10%, label 2 appear 50% and label 3 appear 40% of the training set.
- $\pi_1 = 0.1$, $\pi_2 = 0.5$ and $\pi_3 = 0.4$.

Given a new item with feature value x to classify,

- Compute $\pi_1 P_1(x)$, $\pi_2 P_2(x)$ and $\pi_3 P_3(x)$.
- If $\pi_1 P_1(x)$ is the maximum, label it 1.
- If $\pi_2 P_2(x)$ is the maximum, label it 2.
- If $\pi_3 P_3(x)$ is the maximum, label it 3.

Example:
Data space $\mathcal{X} = \mathbb{R}$
Classes/labels $\mathcal{Y} = \{1, 2, 3\}$



For each class j, we have:

- the probability of that class, $\pi_j = \Pr(y = j)$
- the distribution of data in that class, $P_j(x)$

Overall **joint distribution**: $\Pr(x, y) = \Pr(y)\Pr(x|y) = \pi_y P_y(x)$.

To classify a new x: pick the label y with largest $\Pr(x, y)$

Review: Probability Space, Event

Example 1.

Random Experiment: Roll two dices. What is the probability that they add up to 10?

The probability space for this experiment has two components to it:

1. The sample space: The set of all possible outcomes.

   (1, 1), (1, 2), …, (1, 6), (2, 1), …, (2, 6), (3, 1), …, (6,6).

2. The probabilities of each of these outcomes.

   Each of the outcome is equally likely and they sum up to 1.

   So the probability for each of these outcomes is 1/36.

Event: A subset of the possible outcomes.

Event A consists of all pairs of the form (x , y) such that x + y = 10.

A = {(4, 6), (5, 5), (6, 4)}. P(A) = 3/36 = 1/12.

Review: Multiple events

Example 2.

There are 10 coins. Coins 1, 2, …, 9 are fair coins. They turn head and tail 50% of the time.

The tenth coin is not a fair coin. It turns tail all the time.

Experiment:

Pick a coin at random. Throw four times.

What is the probability that tail turned up all four times?

Event A: coin 10 (or the unfair coin) is chosen.

Event B: All four throws resulted in tails.

Review: Conditional Probability

If we know coin 10 is chosen (That is, the event A happened) we know the probability that all four throws turned up tails (That is, P(B)) is 1.

Thus $P(B|A) = 1$.

If coin 10 is not chosen

$P(B|A') = .5x.5.x.5x.5 = .0625$

Review: Conditional Probability

For two events A and B, the conditional probability

$P(A|B)$ = the probability that A occurs given that B occurred

$P(A \cap B) = P(A)P(B|A)$

What is P(B)?

$$P(B) = P(A).P(B|A) + P(A').P(B|A')$$

$$= 0.1x1 + 0.9x0.0625 = 0.15625$$

VERY IMPORTANT: $P(B) \neq P(B|A) + P(B|A')$

$P(B|A) + P(B|A') = 1 + 0.0625 = 1.0625.$

Bayes' Rule

$P(A \cap B) = P(A).P(B|A)$.
$P(A \cap B) = P(B).P(A|B)$.

$P(B).P(A|B) = P(A).P(B|A)$.

Thus

$P(A|B) = P(A).P(B|A) / P(B)$

$P(A) = 0.1$, $P(B|A) = 1$ and $P(B) = 0.15625$
$P(A|B) = .1/0.15625 = 0.64$

Bayes Rule

Example. Random experiment: Throw two dices.

Event A : Sum of values in two dices is 10

Event B : First dice has value 4

	1	2	3	4	5	6
1	2	3	4	5	6	
2	3	4	5	6		
3	4	5	6			
4	5	6				10
5	6				10	
6				10		

$P(A) = 3/36 = 1/12 \quad P(B) = 6/36 = 1/6 \quad P(A \cap B) = 1/36$

$P(A|B) = 1/6 \quad P(B|A) = 1/3$

$P(A \cap B) = 1/36 = (1/6)(1/6) = P(A|B)\,P(B)$

$P(A \cap B) = 1/36 = (1/3)(1/12) = P(B|A)\,P(A)$

$P(A|B) = P(A).P(B|A) / P(B)$ (Bayes Rule)

$1/6 = (1/12)(1/3)/(1/6)$

Random variable

Roll two dice. Let X be their sum.

$$\text{outcome} = (1,1) \quad \Rightarrow \quad X = 2$$
$$\text{outcome} = (1,2) \text{ or } (2,1) \quad \Rightarrow \quad X = 3$$

Probability space:

- Sample space: $\Omega = \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$.
- Each outcome equally likely.

Random variable X lies in $\{2,3,4,5,6,7,8,9,10,11,12\}$.

A **random variable (r.v.)** is a defined on a probability space.
It is a mapping from Ω (outcomes) to $\mathbb{R}$ (numbers).
We'll use capital letters for r.v.'s.

Quiz

Prove

1. $P(A \mid B) = P(A \cap B) / P(B)$
2. Bayes Rule
3. $P(B) = P(A).P(B|A) + P(A').P(B|A')$
4. $P(B) = P(A_1).P(B|A_1) + \ldots + P(A_n).P(B|A_n)$

   where A_i (I = 1, …, n) is a partition of S, the sample space.

Expected value

Expected value, or mean

Expected value of a random variable X:

$$\mathbb{E}(X) = \sum_x x \Pr(X = x).$$

Roll a die. Let X be the number observed.
What is $\mathbb{E}(X)$?

Expected value

x	P(x)	x.P(x)
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	6/6

$E(X) = 1/6 + 2/6 + 3/6 + 4/6 + 5/6 + 6/6 = 21/6 = 3.5$

E(X) Example 2

A biased coin which turns head with probability p.

Random variable X. x = 1 if head turns up and 0 otherwise. What is E(X)?

E(X) Example 2

A biased coin which turns head with probability p.

Random variable X. x = 1 if head turns up and 0 otherwise. What is E(X)?

$E(X) = 1*p + 0*(1-p) = p.$

E(X) : Example 3

Throw two dices. X is the sum of the values.

Range(X) = {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}

$$E(X) = 2*(1/36) + 3*(2/36) + 4*(3/36) + 5*(4/36) + 6*(5/36) + 7*(6/36) + 8*(5/36) + 9*(4/36) + 10*(3/36) + 11*(2/36) + 12*(1/36)$$

Property of E(X)

Let $Y = aX + b$.

Then $E(Y) = aE(X) + b$

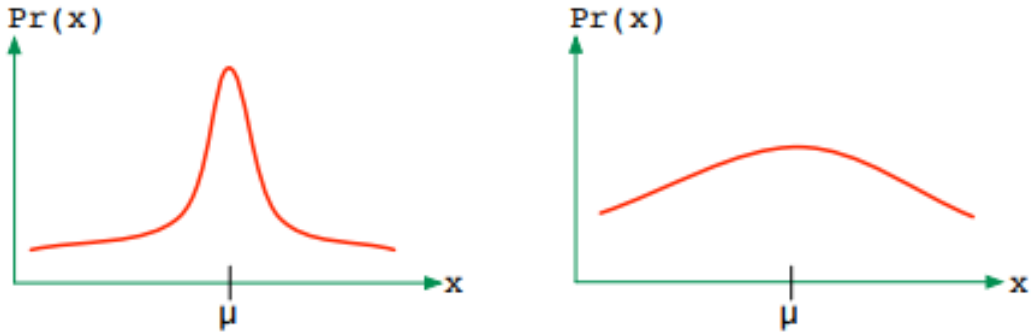
Let f be a polynomial.

$E(f(X)) = f(E(X))$

Is $E(X*X) = E(X)*E(X)$?

Variance

Can summarize an r.v. X by its mean, μ. But this doesn't capture the **spread** of X:



A measure of spread: average distance from the mean, $\mathbb{E}(|X - \mu|)$?

- **Variance:** $\text{var}(X) = \mathbb{E}((X - \mu)^2)$, where $\mu = \mathbb{E}(X)$
- **Standard deviation** $\sqrt{\text{var}(X)}$:
  Roughly, the average amount by which X differs from its mean.

Another formula and property

$$E((X-\mu)^2) = E(X^2 - 2\mu X + \mu^2)$$
$$= E(X^2) - 2\mu E(X) + \mu^2$$
$$= E(X^2) - 2\mu^2 + \mu^2$$
$$= E(X^2) - \mu^2$$
$$= E(X^2) - E(X)^2$$

Let $Y = aX + b$.

Then $\text{Variance}(Y) = a^2\text{Variance}(X)$

Example

A number is chosen from the set {1, 2, 3, 4, 5} at random.

$E(X) = 1(1/5)+2(1/5)+3(1/5)+4(1/5)+5(1/5)$

$= 15/5 = 3.$

$E(X^2) = 1(1/5)+4(1/5)+9(1/5)+16(1/5)+25(1/5)$

$= 55/5 = 11.$

Variance = $E(X^2) – E(X)^2 = 11 – 9 = 2.$

Standard deviation = Sqrt(variance) = Sqrt(2)

Main Point 1

A Bayes classifier starts by calculating the relevant parameters for each class based on the population. Thus the first step is to model the probability distribution function for each class. *All expressions in the universe, however chaotic they may appear, are governed by laws of nature which are grounded in the home of all the laws of nature, the field of pure intelligence.*