Naive Bayes Classifier

Gaussian Event Model

Aditya A Prasad[1]

[1]Department of Instrumentation and Control
National Institute of Technology, Tiruchirappalli

Spider KS, March 12 2017

Outline

Outline

The Iris Data set

Multiclass classification problem

- 3 classes - $[C_1, C_2, C3]$

Figure: Setosa, Virginica, Versicolor

The Iris Data set

Multiclass classification problem

- 4 features - $\vec{x} = [x_1, x_2, x_3, x_4]$

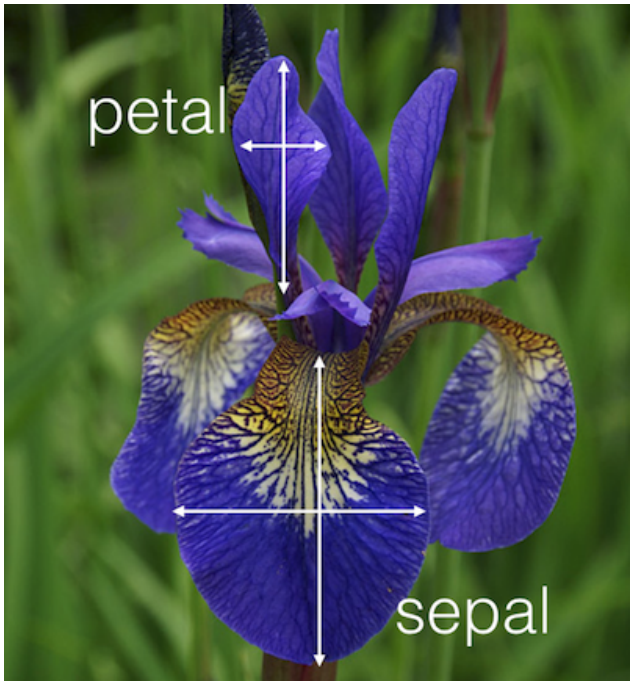

Figure: Sepal and Petal, width and length

Posterior Probability

- $P(C_k|\vec{x}) = ??$

Posterior Probability

- $P(C_k|\vec{x}) = ??$
- We can't calculate this for all possible $\vec{x}$. So we need to reformulate this problem.

Outline

Visualize it

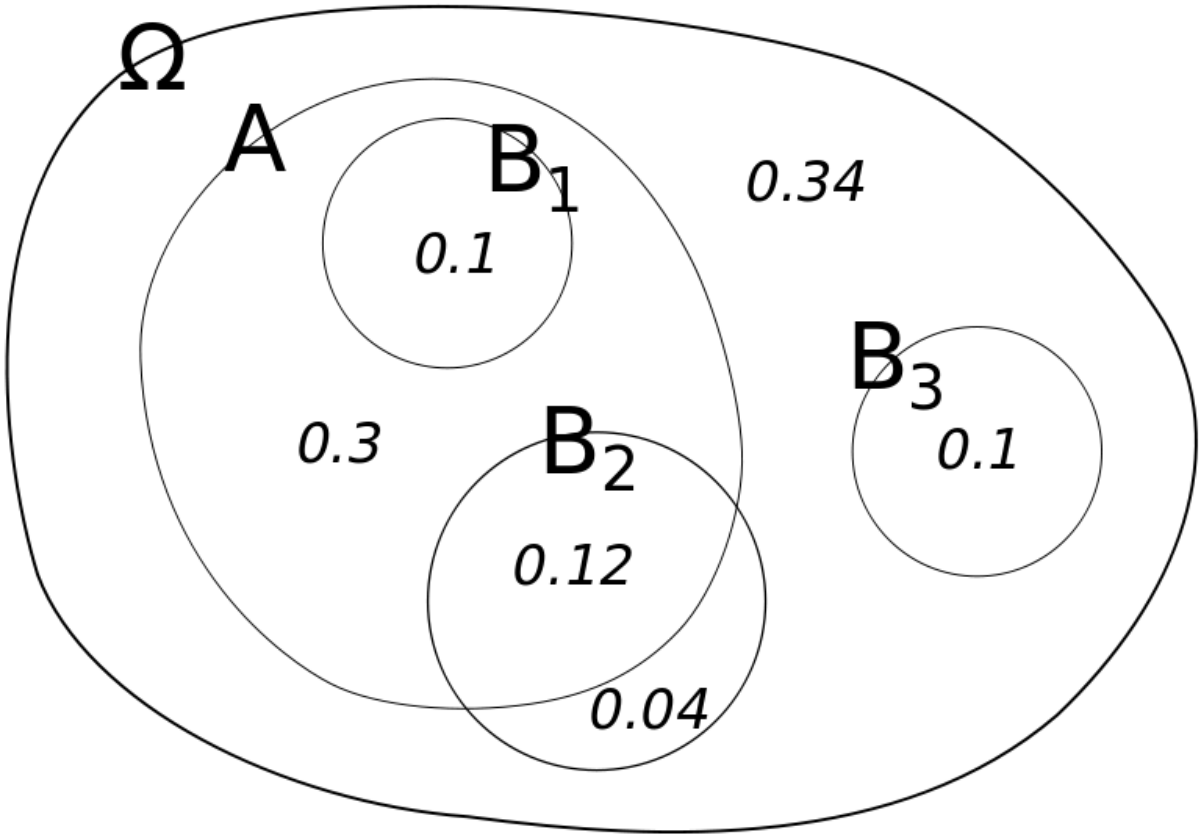

Figure: Venn diagram

Some probability terms

- A sample space of an experiment is the set of all possible outcomes.

Some probability terms

- A sample space of an experiment is the set of all possible outcomes.
- Any subset of the sample space is called an event.

Some probability terms

- A sample space of an experiment is the set of all possible outcomes.
- Any subset of the sample space is called an event.
- $P(A)$ is the probability of observing event A

Some probability terms

- A sample space of an experiment is the set of all possible outcomes.
- Any subset of the sample space is called an event.
- $P(A)$ is the probability of observing event A
- If the event of interest is A and the event B is known or assumed to have occurred, "the conditional probability of A given B", i.e. $P(A|B)$ is the probability of observing event A given that B is *true*.

Some probability terms

- A sample space of an experiment is the set of all possible outcomes.
- Any subset of the sample space is called an event.
- $P(A)$ is the probability of observing event A
- If the event of interest is A and the event B is known or assumed to have occurred, "the conditional probability of A given B", i.e. $P(A|B)$ is the probability of observing event A given that B is *true*.
- **There need not be a causal or temporal relationship between A and B**

Conditional Probability

- $$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Conditional Probability

- $$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

  Restricting the sample space to B

Conditional Probability

- $$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

  Restricting the sample space to B
- If the outcomes are restricted to B, this set now serves as the new sample space.

Conditional Probability

- $$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

  Restricting the sample space to B
- If the outcomes are restricted to B, this set now serves as the new sample space.
- $$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Conditional Probability

- $$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

  Restricting the sample space to B
- If the outcomes are restricted to B, this set now serves as the new sample space.
- $$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

  Restricting the sample space to A

Visualize it

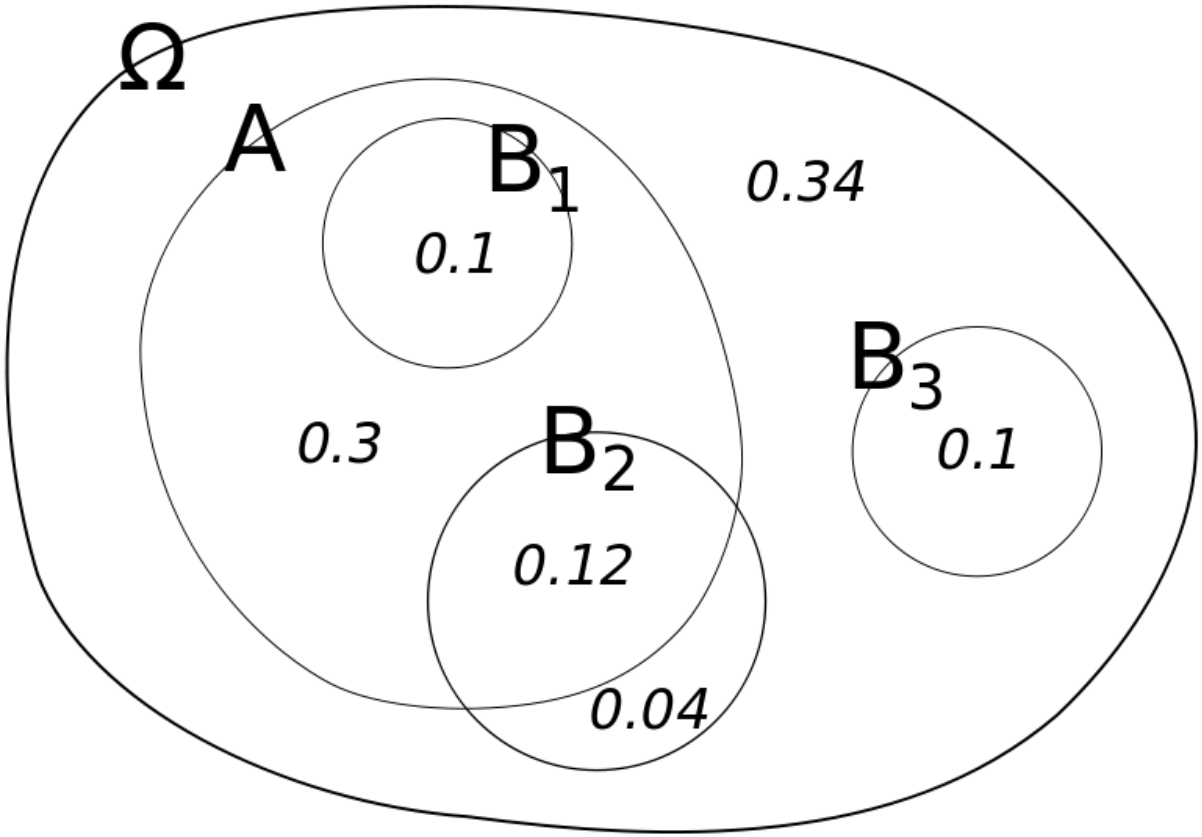

Figure: Venn diagram

Bayes Theorem

- Joint probability distribution

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

Bayes Theorem

- Joint probability distribution

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

- Thus we have the famed equation,

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Bayes Theorem

- Joint probability distribution

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

- Thus we have the famed equation,

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Bayes Theorem

- Joint probability distribution

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

- Thus we have the famed equation,

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

- Or in english

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

Bayes Theorem

- Joint probability distribution

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

- Thus we have the famed equation,

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

- Or in english

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

Note how we update our beliefs faced with evidence

Coming to our flowers

- So using Bayes theorem we can write,

$$P(C_k|\vec{x}) = \frac{P(C_k)P(\vec{x}|C_k)}{P(\vec{x})}$$

Coming to our flowers

- So using Bayes theorem we can write,

$$P(C_k|\vec{x}) = \frac{P(C_k)P(\vec{x}|C_k)}{P(\vec{x})}$$

- We need to compare $P(C_1, \vec{x}), P(C_2, \vec{x}), P(C_3, \vec{x})$

Coming to our flowers

- So using Bayes theorem we can write,

$$P(C_k|\vec{x}) = \frac{P(C_k)P(\vec{x}|C_k)}{P(\vec{x})}$$

- We need to compare $P(C_1, \vec{x}), P(C_2, \vec{x}), P(C_3, \vec{x})$

Coming to our flowers

- So using Bayes theorem we can write,

$$P(C_k|\vec{x}) = \frac{P(C_k)P(\vec{x}|C_k)}{P(\vec{x})}$$

- We need to compare $P(C_1, \vec{x}), P(C_2, \vec{x}), P(C_3, \vec{x})$
- So only the likelihood term matters.

$$C^* = \textbf{argmax}_{C_k} P(\vec{x}|C_k)$$

Multivariate modelling

- For each class we would need to try to fit a model - like the Gaussian multivariate.

Multivariate modelling

- For each class we would need to try to fit a model - like the Gaussian multivariate.
- Curse of dimensionality - the volume of the space increases so fast that the available data become sparse.

Multivariate modelling

- For each class we would need to try to fit a model - like the Gaussian multivariate.
- Curse of dimensionality - the volume of the space increases so fast that the available data become sparse.
  Harder to find groups with similar properties

Multivariate modelling

- For each class we would need to try to fit a model - like the Gaussian multivariate.
- Curse of dimensionality - the volume of the space increases so fast that the available data become sparse.
  Harder to find groups with similar properties
  Not all combinations are seen

Multivariate modelling

- For each class we would need to try to fit a model - like the Gaussian multivariate.
- Curse of dimensionality - the volume of the space increases so fast that the available data become sparse.
  Harder to find groups with similar properties
  Not all combinations are seen
  Distance metrics lose meaning
  and so on ....

Multivariate modelling

- For each class we would need to try to fit a model - like the Gaussian multivariate.
- Curse of dimensionality - the volume of the space increases so fast that the available data become sparse.
  Harder to find groups with similar properties
  Not all combinations are seen
  Distance metrics lose meaning
  and so on ....
- A Naive assumption can turn this into modelling each feature separately i.e. n univariate distributions.

Outline

Not Independence

- A and B are conditionally independent given Q if and only if, **given knowledge that Q occurs**, knowledge of whether A occurs provides no information on the likelihood of B occurring, and knowledge of whether B occurs provides no information on the likelihood of A occurring.

Not Independence

- A and B are conditionally independent given Q if and only if, **given knowledge that Q occurs**, knowledge of whether A occurs provides no information on the likelihood of B occurring, and knowledge of whether B occurs provides no information on the likelihood of A occurring.
- Equation form this would be,

$$P(A \cap B \mid Q) = P(A \mid Q)P(B \mid Q)$$

or equivalently,

$$P(A \mid B \cap Q) = P(A \mid Q)$$

Not Independence

- A and B are conditionally independent given Q if and only if, **given knowledge that Q occurs**, knowledge of whether A occurs provides no information on the likelihood of B occurring, and knowledge of whether B occurs provides no information on the likelihood of A occurring.
- Equation form this would be,

$$P(A \cap B \mid Q) = P(A \mid Q)P(B \mid Q)$$

or equivalently,

$$P(A \mid B \cap Q) = P(A \mid Q)$$

- Height and vocabulary are not independent; but they are conditionally independent if you add age.

Not Independence

- A and B are conditionally independent given Q if and only if, **given knowledge that Q occurs**, knowledge of whether A occurs provides no information on the likelihood of B occurring, and knowledge of whether B occurs provides no information on the likelihood of A occurring.
- Equation form this would be,

$$P(A \cap B \mid Q) = P(A \mid Q)P(B \mid Q)$$

or equivalently,

$$P(A \mid B \cap Q) = P(A \mid Q)$$

- Height and vocabulary are not independent; but they are conditionally independent if you add age. Two dice throws are independent; if P was sum of the two results is even then they **are not** conditionally independent w.r.t P

Joint Probability

- We can rewrite $P(A \cap B) = P(A)P(B|A) = P(A, B)$ as,

$$P(C_k, x_1, x_2, ...x_n) = P(x_1|x_2, x_3, ...x_n, C_k)P(x_2, x_3, ...x_n)$$

$$= P(x_1|x_2, x_3, ...x_n, C_k)P(x_2|x_3, ...x_n)P(x_3|x_4, ..., x_n)...P(x_n|C_k)P(C_k)$$

Using chain rule of conditional probability.

Joint Probability

- We can rewrite $P(A \cap B) = P(A)P(B|A) = P(A, B)$ as,

$$P(C_k, x_1, x_2, ...x_n) = P(x_1|x_2, x_3, ...x_n, C_k)P(x_2, x_3, ...x_n)$$

$$= P(x_1|x_2, x_3, ...x_n, C_k)P(x_2|x_3, ...x_n)P(x_3|x_4, ..., x_n)...P(x_n|C_k)P(C_k)$$

Using chain rule of conditional probability.

- Note how an assumption of conditional independence between any two features given the class would simplify this to,

$$P(C_k, \vec{x}) = P(x_1|C_k)P(x_2|C_k)...P(x_n|C_k)P(C_k) = P(C_k)\prod_{i=1}^{n} P(x_i|C_k)$$

The naive Bayes probability model

- $$P(C_k)P(\vec{x}|C_k) = P(C_k, \vec{x}) = P(C_k)\prod_{i=1}^{n} P(x_i|C_k)$$

The naive Bayes probability model

- $$P(C_k)P(\vec{x}|C_k) = P(C_k, \vec{x}) = P(C_k)\prod_{i=1}^{n} P(x_i|C_k)$$

- $$P(\vec{x}|C_k) = \prod_{i=1}^{n} P(x_i|C_k)$$

Outline

Blocks

Block Title

You can also highlight sections of your presentation in a block, with it's own title

Theorem

There are separate environments for theorems, examples, definitions and proofs.

Example

Here is an example of an example block.

Summary

- The first main message of your talk in one or two lines.
- The second main message of your talk in one or two lines.
- Perhaps a third message, but not more than that.

- Outlook
  - Something you haven't solved.
  - Something else you haven't solved.

For Further Reading I

A. Author.
Handbook of Everything.
Some Press, 1990.

S. Someone.
On this and that.
Journal of This and That, 2(1):50–100, 2000.