Representing Uncertainty + Probabilistic Learning

R&N Chapter 13

A bit of 20.2

Uncertainty

- Most real-world problems deal with uncertain information
  - Diagnosis: Likely disease given observed symptoms
  - Equipment repair: Likely component failure given sensor reading
  - Help desk: Likely operation based on past operations
  - Cannot be represented by deterministic rules

Headache => Fever

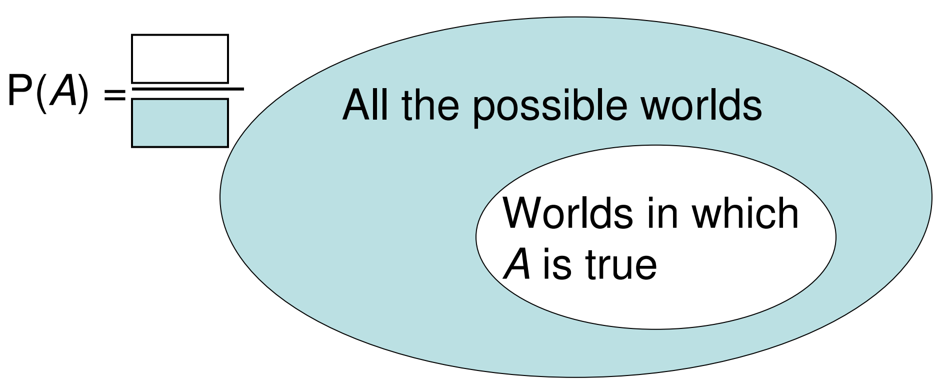
Uncertainty

- Correct framework for representing uncertainty: Probability
- Outline:
  - Review of basic probability tools (much of it well-known, but still important to review)
  - Bayes rule and its use in uncertain reasoning and probabilistic learning

Probability

- P(*A*) = Probability of event *A* = percentage of all possible worlds in which *A* is true.

$$0 \leq \mathrm{P}(A) \leq 1$$



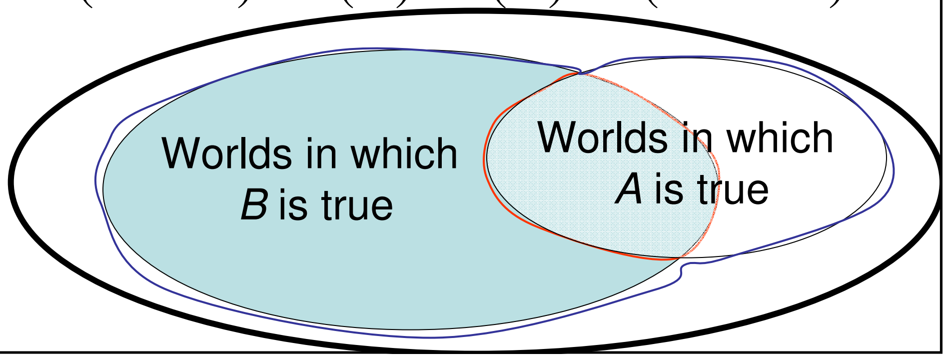
Probability

$$0 \leq P(\boldsymbol{A}) \leq 1$$

$$P(\boldsymbol{True}) = 1$$

$$P(\boldsymbol{False}) = 0$$

$$P(\boldsymbol{A} \text{ or } \boldsymbol{B}) = P(\boldsymbol{A}) + P(\boldsymbol{B}) - P(\boldsymbol{A} \text{ and } \boldsymbol{B})$$



Probability

$$0 \leq P(\boldsymbol{A}) \leq 1$$

$$P(\boldsymbol{True}) = 1$$

$$P(\boldsymbol{False}) = 0$$

$$P(\boldsymbol{A} \text{ or } \boldsymbol{B}) = P(\boldsymbol{A}) + P(\boldsymbol{B}) - P(\boldsymbol{A} \text{ and } \boldsymbol{B})$$

- Other ideas:
  - Fuzzy logic
  - Non-monotonic logic
  - Multi-valued logic
  - Evidence theory (Dempster-Shafer)
- Probability is the only system that is “consistent”

Probability

- Immediately derived properties

$$P(\neg A) = 1 - P(A)$$

Denotes not-*A* = All the worlds in which *A* does not occur

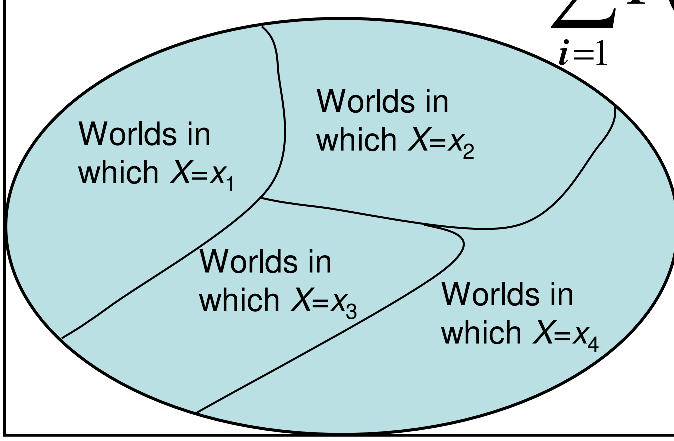
$$P(A) = P(A,B) + P(A,\neg B)$$

Short hand for "*A* and *B*"

Probability

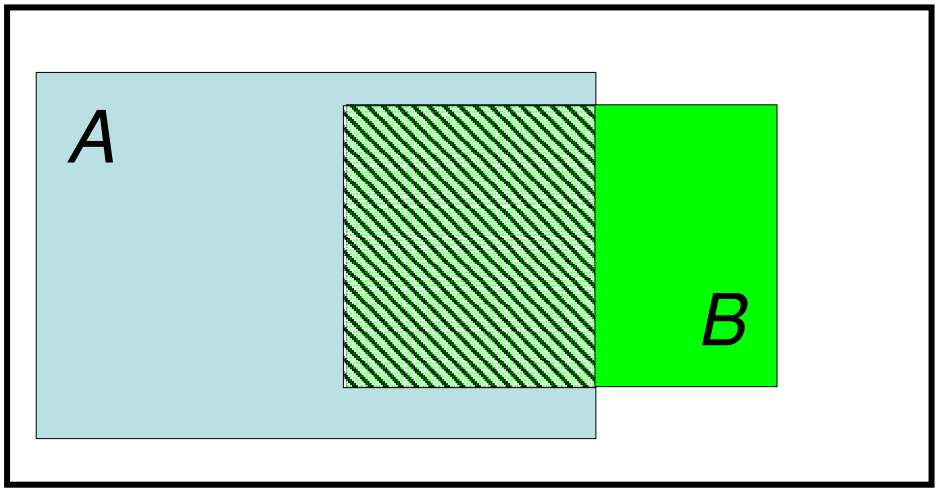
- A random variable is a variable *X* that can take values $x_1,..,x_n$ with a probability $P(X = x_i)$ attached to each $i = 1,..,n$

$$\sum_{i=1}^{n} P(X = x_i) = 1$$



Conditional Probability

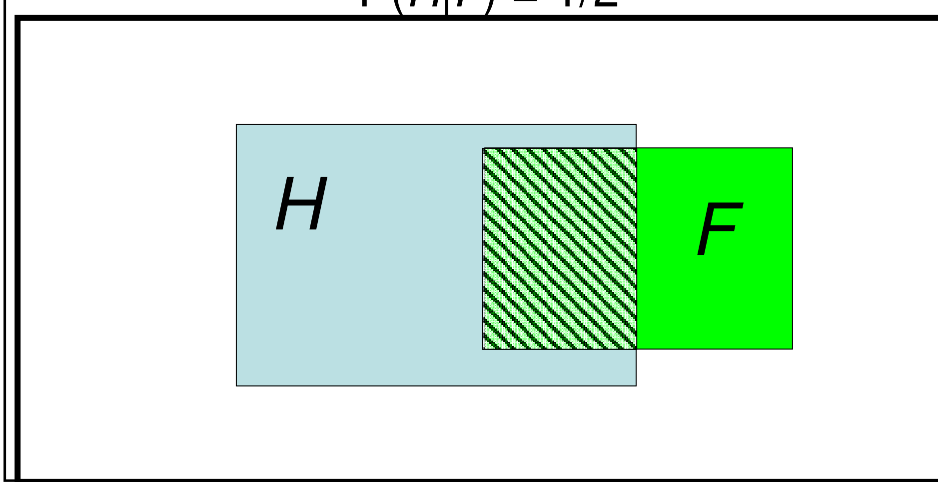
- P(*A*|*B*) = Fraction *of those worlds in which* B *is true* for which *A* is also true.



Conditional Probability Example

- *H* = Headache P(*H*) = 1/2
- *F* = Flu P(*F*) = 1/8

P(*H*|*F*) = 1/2



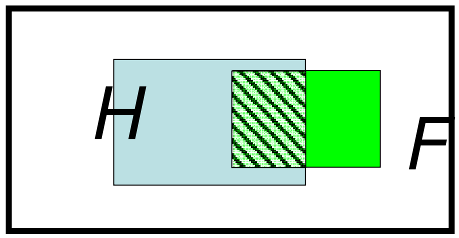
Conditional Probability Example

- H = Headache　　　　P(H) = 1/2
- F = Flu　　　　P(F) = 1/8

$$P(H|F) = \frac{\text{(Area of “}H\text{ and }F\text{” region)}}{\text{(Area of }F\text{ region)}}$$

P(H|F) = P(H,F)/P(F)

P(H|F) = 1/2



Conditional Probability

- Definition:

$$P(\boldsymbol{A} \mid \boldsymbol{B}) = \frac{P(\boldsymbol{A},\boldsymbol{B})}{P(\boldsymbol{B})}$$

- Chain rule:

$$P(\boldsymbol{A},\boldsymbol{B}) = P(\boldsymbol{A} \mid \boldsymbol{B})\,P(\boldsymbol{B})$$

Conditional Probability

- Other useful relations:

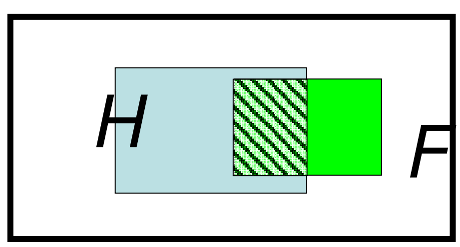
$$P(A \mid B) + P(\neg A \mid B) = 1$$

$$\sum_i P(X = x_i \mid B) = 1$$

Probabilistic Inference

- Suppose H is true
- Suppose that you know P(*H*|*F*) = ½ = 0.5
- What is the probability that *F* is true? 0.5?

- P(*H*) = 1/2
- P(*F*) = 1/8
- P(*H*|*F*) = 0.5



Probabilistic Inference

- Correct reasoning:
- We know P(*H*), P(*F*), P(*H*|*F*) and the two chain rules:

$$\mathrm{P}(\boldsymbol{H},\boldsymbol{F}) = \mathrm{P}(\boldsymbol{H} \mid \boldsymbol{F})\,\mathrm{P}(\boldsymbol{F})$$

$$\mathrm{P}(\boldsymbol{F} \mid \boldsymbol{H}) = \frac{\mathrm{P}(\boldsymbol{H},\boldsymbol{F})}{\mathrm{P}(\boldsymbol{H})}$$

- Substituting the values:

$$\mathrm{P}(\boldsymbol{H},\boldsymbol{F}) = 0.5 \times 1/8 = 1/16$$

$$\mathrm{P}(\boldsymbol{F} \mid \boldsymbol{H}) = \frac{1/16}{1/2} = \boxed{1/8}$$

Probabilistic Inference

- Correct reasoning:
- We know P(*H*), P(*F*), P(*H*|*F*) and the two chain rules:

$$\mathrm{P}(\boldsymbol{H},\boldsymbol{F}) = \mathrm{P}(\boldsymbol{H} \mid \boldsymbol{F})\,\mathrm{P}(\boldsymbol{F})$$

$$\mathrm{P}(\boldsymbol{F} \mid \boldsymbol{H}) = \frac{\mathrm{P}(\boldsymbol{H},\boldsymbol{F})}{\mathrm{P}(\boldsymbol{H}}$$

- Substituting the values:

$$\mathrm{P}(\boldsymbol{H},\boldsymbol{F}) = 0.5 \times 1/$$

$$\mathrm{P}(\boldsymbol{F} \mid \boldsymbol{H}) = \frac{1/16}{1/2} =$$

The key difference is that we took into account the fact that catching the flu is unlikely (P(*F*) is small)

Bayes Rule

$$\boxed{\mathrm{P}(\boldsymbol{B} \mid \boldsymbol{A}) = \frac{\mathrm{P}(\boldsymbol{A} \mid \boldsymbol{B})\,\mathrm{P}(\boldsymbol{B})}{\mathrm{P}(\boldsymbol{A})}}$$

Introduced circa 1763

Probabilistic Inference

We want: *Posterior* probability that *B* occurs given that *A* occurs

We know: *Prior* probability that *B* occurs in the absence of any other information

$$\mathrm{P}(\boldsymbol{B} \mid \boldsymbol{A}) = \frac{\mathrm{P}(\boldsymbol{A} \mid \boldsymbol{B})\,\mathrm{P}(\boldsymbol{B})}{\mathrm{P}(\boldsymbol{A})}$$

We know: *Likelihood* that *A* occurs given that *B* occurs

Bayes Rule

- What if we do not know P(*A*)???
- Use the relation:

$$\mathrm{P}(\boldsymbol{A}) = \mathrm{P}(\boldsymbol{A} \mid \boldsymbol{B})\,\mathrm{P}(\boldsymbol{B}) + \mathrm{P}(\boldsymbol{A} \mid \neg\boldsymbol{B})\,\mathrm{P}(\neg\boldsymbol{B})$$

- More general Bayes rule:

$$\mathrm{P}(\boldsymbol{B} \mid \boldsymbol{A}) = \frac{\mathrm{P}(\boldsymbol{A} \mid \boldsymbol{B})\,\mathrm{P}(\boldsymbol{B})}{\mathrm{P}(\boldsymbol{A} \mid \boldsymbol{B})\,\mathrm{P}(\boldsymbol{B}) + \mathrm{P}(\boldsymbol{A} \mid \neg\boldsymbol{B})\,\mathrm{P}(\neg\boldsymbol{B})}$$

Bayes Rule

- Same rule for a non-binary random variable, except we need to sum over all the possible events $X = x_i$

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}_i \mid \boldsymbol{A}) = \frac{\mathrm{P}(\boldsymbol{A} \mid \boldsymbol{X} = \boldsymbol{x}_i)\,\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}_i)}{\mathrm{P}(\boldsymbol{A})}$$

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}_i \mid \boldsymbol{A}) = \frac{\mathrm{P}(\boldsymbol{A} \mid \boldsymbol{X} = \boldsymbol{x}_i)\,\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}_i)}{\sum_k \mathrm{P}(\boldsymbol{A} \mid \boldsymbol{X} = \boldsymbol{x}_k)\,\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}_k)}$$

This is actually just P(*A*)

Joint Distribution

- Joint Distribution Table:
  - Given a set of variables A,B,C,….
  - Generate a table with all the possible combinations of assignments to the variables in the rows
  - For each row, list the corresponding joint probability
  - For M binary variables → size 2^M

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

Using the Joint Distribution: Computing Other Probabilities

Compute the probability of event *E*:

$$\boldsymbol{P}(\boldsymbol{E}) = \sum_{\substack{\text{all rows} \\ \text{containing } \boldsymbol{E}}} \boldsymbol{P}(\text{row})$$

$\boldsymbol{P}(\boldsymbol{A},\boldsymbol{B}) =$ 0.25 + 0.10 = 0.35

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

Using the Joint Distribution: Doing Inference

Given that event E_1 occurs, what is the probability that E_2 occurs:

$$P(E_2 \mid E_1) = \frac{P(E_2, E_1)}{P(E_1)}$$

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

Using the Joint Distribution: Doing Inference

$$P(A, B \mid C) = \frac{P(A, B, C)}{P(C)}$$

$$= \frac{0.10}{0.05+0.05+0.10+0.10} = \frac{0.10}{0.30}$$

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

Inference

- General view: I have some evidence (Headache) how likely is a particular conclusion (Fever)
- Important in many industries: Medical, pharmaceutical, Help Desk, Fault Diagniosis….

Learning the Joint Distribution

- Three possible ways of generating the joint distribution:
  1. Human experts (very difficult!)
  2. Using known conditionally probabilities (e.g., if we know P(C|A,B), P(B|A), and P(A), we know P(A,B,C) = P(C|A,B)P(B|A)P(A) → This is the basis for *Bayes Nets*, to be covered later….)
  3. *Learning from data*

Learning the Joint Distribution

Suppose that we have recorded a lot of training data:

(0,1,1)
(1,0,1)
(1,1,0)
(0,0,0)
(1,1,0)...........

The entry for P(A,B,¬C) in the table is:

$$\frac{\text{\# of data entries with A=1, B=1, C=0}}{\text{Total number of data entries}}$$

A	B	C	Prob
0	0	0	?
0	0	1	?
0	1	0	?
0	1	1	?
1	0	0	?
1	0	1	?
1	1	0	?
1	1	1	?

Learning the Joint Distribution

Suppose that we have recorded a lot of training data:

(0,1,1)
(1,0,1)
(1,1,0)
(0,0,0)
(1,1,0)...........

More generally, the entry for P(E) in the table is:

$$\frac{\text{\# of data entries with E}}{\text{Total number of data entries}}$$

A	B	C	Prob
0	0	0	?
0	0	1	?
0	1	0	?
0	1	1	?
1	0	0	?
1	0	1	?
1	1	0	?
1	1	1	?

Real-Life Joint Distribution

- UCI Census Database

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Real-Life Joint Distribution

- UCI Census Database

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Male|Poor) = 0.4654/0.7604 = 0.612

So Far….

- Basic probability concepts
- Bayes rule
- What are joint distributions
- Inference using joint distributions
- Learning joint distributions from data

- Problem: If we have *M* variables, we need 2^M entries in the joint distribution table → An independence assumption leads to an efficient way to learn and to do inference

Independence

- *A* and *B* are independent iff:

$$\boxed{P(A \mid B) = P(A)}$$

- In words: Knowing *B* does not affect how likely we think that *A* is true

Key Properties

- Symmetry:

$$\mathrm{P}(\boldsymbol{A} \mid \boldsymbol{B}) = \mathrm{P}(\boldsymbol{A}) \Leftrightarrow \mathrm{P}(\boldsymbol{B} \mid \boldsymbol{A}) = \mathrm{P}(\boldsymbol{B})$$

- Joint distribution:

$$\mathrm{P}(\boldsymbol{A}, \boldsymbol{B}) = \mathrm{P}(\boldsymbol{A})\,\mathrm{P}(\boldsymbol{B})$$

- Independence of complements:

$$\mathrm{P}(\neg \boldsymbol{A} \mid \boldsymbol{B}) = \mathrm{P}(\neg \boldsymbol{A}) \qquad \mathrm{P}(\boldsymbol{A} \mid \neg \boldsymbol{B}) = \mathrm{P}(\boldsymbol{A})$$

Naïve Bayes

- Suppose that *A*, *B*, *C* are *independent*
- Then any value of the joint distribution can be computed easily:

$$\mathrm{P}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = \mathrm{P}(\boldsymbol{A})\,\mathrm{P}(\boldsymbol{B})\,\mathrm{P}(\boldsymbol{C})$$

$$\mathrm{P}(\boldsymbol{A}, \neg \boldsymbol{B}, \boldsymbol{C}) = \mathrm{P}(\boldsymbol{A})\,\mathrm{P}(\neg \boldsymbol{B})\,\mathrm{P}(\boldsymbol{C})$$

- In fact, we need only *M* numbers instead of 2^M for binary variables!!

Naïve Bayes: General Case

- If $X_1,..,X_M$ are independent variables:

$$\mathrm{P}(\boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2, \ldots, \boldsymbol{X}_M = \boldsymbol{x}_M) =$$
$$\mathrm{P}(\boldsymbol{X}_1 = \boldsymbol{x}_1)\,\mathrm{P}(\boldsymbol{X}_2 = \boldsymbol{x}_2)\ldots\mathrm{P}(\boldsymbol{X}_M = \boldsymbol{x}_M)$$

- Under the "Naïve" assumption, we can compute any value of the joint distribution
- We can answer any inference query
- How do we learn the distributions?

Naïve Bayes: Learning

$$\boxed{\boldsymbol{P}(\boldsymbol{X}_i = \boldsymbol{x}) = \frac{\text{Number of observations with } \boldsymbol{X}_i = \boldsymbol{x}}{\text{Total Number of observations}}}$$

- Learning the distributions from data is simple and efficient
- In practice, the independence assumption may not be met but it is often a very useful approximation (see examples at the end)

So Far….

- Basic probability concepts
- Bayes rule
- What are joint distributions
- Inference using joint distributions
- Learning joint distributions from data
- Independence assumption
- Naïve Bayes

- Problem: We now have the joint distribution. How can we use it to make decision → Bayes Classifier

Problem Example

- Three variables:
  - Hair = {blond, dark}
  - Height = {tall,short}
  - Country = {Gromland, Polvia}
- Training data: Values of (Eye,Height,Country) collected over population

Joint Distribution Table:

(B,T,G)	(B,T,P)
(D,T,G)	(B,T,P)
(D,T,G)	(B,T,P)
(D,T,G)	(D,T,P)
(B,T,G)	(D,T,P)
(B,S,G)	(D,S,P)
(B,S,G)	(B,S,P)
(D,S,G)	(D,S,P)

Learn Joint Probabilities

- Three variables:
  - Hair = {blond, dark}
  - Height = {tall,short}
  - Country = {Gromland, Polvia}
- Training data: Values of (Eye,Height,Country) collected over population

(B,T,G)	(B,T,P)	P(B,S,G)= 2/16
(D,T,G)	(B,T,P)	P(B,T,G)= 2/16
(D,T,G)	(B,T,P)	P(D,S,G)= 1/16
(D,T,G)	(D,T,P)	P(D,T,G)= 3/16
(B,T,G)	(D,T,P)	P(B,S,P)= 1/16
(B,S,G)	(D,S,P)	P(B,T,P)= 3/16
(B,S,G)	(B,S,P)	P(D,S,P)= 2/16
(D,S,G)	(D,S,P)	P(D,T,P)= 2/16

Compute other Joint Or Conditional Distributions

P(B,S,G)= 2/16
P(B,T,G)= 2/16
P(D,S,G)= 1/16
P(D,T,G)= 3/16
P(B,S,P)= 1/16
P(B,T,P)= 3/16
P(D,S,P)= 2/16
P(D,T,P)= 2/16

$$P(\text{Hair} = B, \text{Height} = S | \text{Country}=G) = \frac{P(\text{Hair} = B, \text{Height} = S, \text{Country}=G)}{P(\text{Country}=G)} = \frac{2/16}{1/2} = 4/16$$

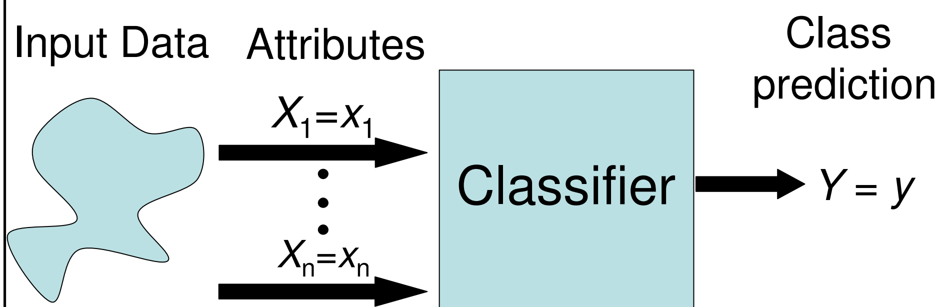
Classifier Example

- Three variables:
  - Hair = {blond, dark}
  - Height = {tall,short}
  - Country = {Gromland, Polvia}
- Training data: Values of (Eye,Height,Country) collected over population

(B,T,G)	(B,T,P)
(D,T,G)	(B,T,P)
(D,T,G)	(B,T,P)
(D,T,G)	(D,T,P)
(B,T,G)	(D,T,P)
(B,S,G)	(D,S,P)
(B,S,G)	(B,S,P)
(D,S,G)	(D,S,P)

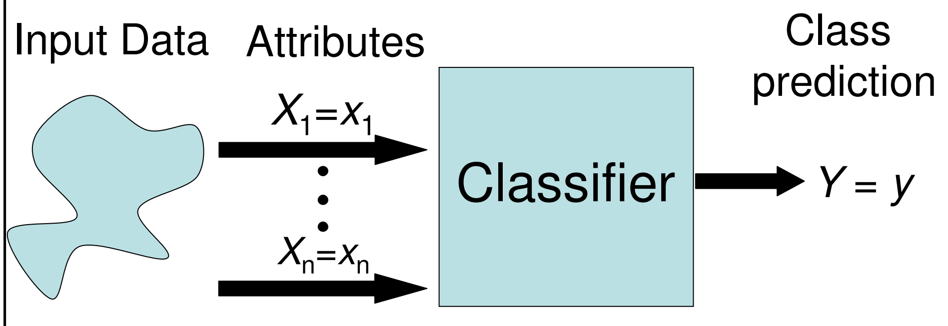
If I observe a new individual tall with blond hair, what is the most likely country of origin?

Classifiers



- How to recover from which class the input data comes?

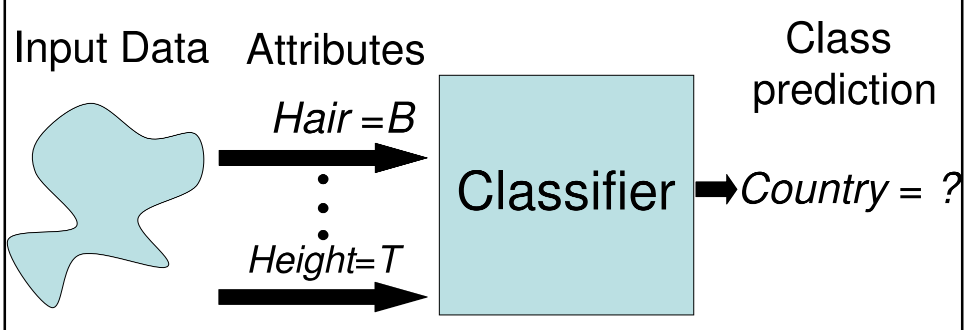
Classifiers



- Ho nput da

Y is called a "class" label here for presentation convenience. In fact, it could be any attribute that can be predicted from the data.

Classifiers



- How to recover from which class the input data comes?

Classifiers

- We want to find the value of *Y* that is the most probable, given the observations $X_1,..,X_n$
- Find *y* such that this is maximum:

$$P(Y = y \mid X_1 = x_1, \ldots, X_n = x_n)$$

Classifiers

- We want to find the value of *Y* that is the most probable, given the observations $X_1,..,X_n$
- Find *y* such that this is maximum:

$$P(Y = y \mid X_1 = x_1, \ldots, X_n = x_n)$$

The maximum is called the *Maximum A Posteriori (MAP)* estimator

Classifiers

- We want to find the value of *Y* that is the most probable, given the observations $X_1,..,X_n$
- Find *y* such that this is maximum:

$$P(Y = y \mid X_1 = x_1, \ldots, X_n = x_n) = \frac{P(X_1 = x_1, \ldots, X_n = x_n \mid Y = y)\, P(Y = y)}{P(X_1 = x_1, \ldots, X_n = x_n)}$$

Classifiers

- We want to find the value of *Y* that is the most probable, given the obs $X_1,..,X_n$
- Find *y* such that this is maximum:

Apply Bayes rule

$$P(Y = y \mid X_1 = x_1, \ldots, X_n = x_n) = \frac{P(X_1 = x_1, \ldots, X_n = x_n \mid Y = y)\, P(Y = y)}{P(X_1 = x_1, \ldots, X_n = x_n)}$$

This denominator does not depend on *y*. It is a constant (as far as *y* is concerned) and can be ignored.

Bayes Classifier

- We want to find the value of Y that is the most probable, given the observations $X_1,..,X_n$
- Find y such that this is maximum:

$$\boxed{\mathrm{P}(\boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n \mid \boldsymbol{Y} = \boldsymbol{y})\,\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y})}$$

Bayes Classifier

- We want to find the value of Y that is the most probable, given the observations $X_1,..,X_n$
- Find y such that this is maximum:

$$\boxed{\mathrm{P}(\boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n \mid \boldsymbol{Y} = \boldsymbol{y})\,\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y})}$$

Likelihood of observing $(x_1,..,x_n)$ from data of class y. This is learned from training data

Probability of each class, also learned from training data

Bayes Classifier

- Learning:
  - Collect all the observations (x_1,..,x_n) for each class y and estimate:

$$\mathrm{P}(\boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n \mid \boldsymbol{Y} = \boldsymbol{y}) = \frac{\#\,\text{observations with } (\boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n) \text{ in class } \boldsymbol{y}}{\text{Total Number of observations in class } \boldsymbol{y}}$$

$$\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y}) = \frac{\#\,\text{observations in class } \boldsymbol{y}}{\text{Total Number of observations}}$$

- Classification:
  - Given a new input (x_1,..,x_n), compute the best class:

$$\boldsymbol{y}^{best} = \arg\max_{y} \mathrm{P}(\boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n \mid \boldsymbol{Y} = \boldsymbol{y})\,\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y})$$

Classifier Example

- Three variables:
  - Hair = {blond, dark}
  - Height = {tall,short}
  - Country = {Gromland, Polvia
- Training data: Values of (E
  collected over population

If I observe a new individual tall with blond hair, what is the most likely country of origin?

(B,T,G)	(B,T,P)
(D,T,G)	(B,T,P)
(D,T,G)	(B,T,P)
(D,T,G)	(D,T,P)
(B,T,G)	(D,T,P)
(B,S,G)	(D,S,P)
(B,S,G)	(B,S,P)
(D,S,G)	(D,S,P)

P(B,T|G)P(G) = 2/8 x 1/2 = 2/16
P(B,T|P)P(P) = 3/8 x 1/2 = 3/16

Conclusion: Country = P

Classifier Example

- Three variables:
  - Hair = {blond, dark}
  - Height = {tall,short}
  - Country = {Gromland, Polvia
- Training data: Values of (E
  collected over population

If I observe a new individual tall with blond hair, what is the most likely country of origin?

(B,T,G)	(B,T,G)	(B,T,P)
(D,T,G)	(D,T,G)	(B,T,P)
(D,T,G)	(D,T,G)	(B,T,P)
(D,T,G)	(D,T,G)	(D,T,P)
(B,T,G)	(B,T,G)	(D,T,P)
(B,S,G)	(B,S,G)	(D,S,P)
(B,S,G)	(B,S,G)	(B,S,P)
(D,S,G)	(D,S,G)	(D,S,P)

$P(B,T|G)P(G) = 2/8 \times 2/3 = 4/24$

$P(B,T|P)P(P) = 3/8 \times 1/3 = 3/24$

Conclusion: Country = G

Classifier Example

- Three
  - Hair
  - Heig
  - Cou
- Trainir
  collect

Note the different conclusion! That's where the "Bayes" part plays a role. We correctly took into account the fact that Gromland is now twice as likely, irrespective of the observation. $P(G) = 2/3$ $P(P) = 1/3$

(B,T,G)	(B,T,G)	(B,T,P)
(D,T,G)	(D,T,G)	(B,T,P)
(D,T,G)	(D,T,G)	(B,T,P)
(D,T,G)	(D,T,G)	(D,T,P)
(B,T,G)	(B,T,G)	(D,T,P)
(B,S,G)	(B,S,G)	(D,S,P)
(B,S,G)	(B,S,G)	(B,S,P)
(D,S,G)	(D,S,G)	(D,S,P)

P(B,T|G)P(G) = /3 = 4/24

P(B,T|P)P(P) = 3/ /3 = 3/24

Conclusion: Country = G

Naïve Bayes Classifier

- Learning: Collect all the observations (x_1,..,x_n) for each class y and estimate:

$$\mathrm{P}(\boldsymbol{X}_i = \boldsymbol{x}_i \mid \boldsymbol{Y} = \boldsymbol{y}) = \frac{\text{Number of observations with } \boldsymbol{X}_i = \boldsymbol{x}_i \text{ in class } \boldsymbol{y}}{\text{Total Number of observations in class } \boldsymbol{y}}$$

$$\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y}) = \frac{\text{Number of observations in class } \boldsymbol{y}}{\text{Total Number of observations}}$$

- Classification:

$$\boldsymbol{y}^{best} = \arg\max_{y} \mathrm{P}(\boldsymbol{X}_1 = \boldsymbol{x}_1 \mid \boldsymbol{Y} = \boldsymbol{y}) \ldots \mathrm{P}(\boldsymbol{X}_n = \boldsymbol{x}_n \mid \boldsymbol{Y} = \boldsymbol{y}) \, \mathrm{P}(\boldsymbol{Y} = \boldsymbol{y})$$

- Learn... for ea...

$$\mathrm{P}(\boldsymbol{X}_i = \boldsymbol{x}_i \mid \boldsymbol{Y} = \boldsymbol{y}) = \frac{\text{Number of observations with } \boldsymbol{X}_i = \boldsymbol{x}_i \text{ in class } \boldsymbol{y}}{\text{Total Number of observations in class } \boldsymbol{y}}$$

$$\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y}) = \frac{\text{Number of observations in class } \boldsymbol{y}}{\text{Total Number of observations}}$$

- Classification:

$$\boldsymbol{y}^{best} = \arg\max_{y} \mathrm{P}(\boldsymbol{X}_1 = \boldsymbol{x}_1 \mid \boldsymbol{Y} = \boldsymbol{y}) \ldots \mathrm{P}(\boldsymbol{X}_n = \boldsymbol{x}_n \mid \boldsymbol{Y} = \boldsymbol{y}) \, \mathrm{P}(\boldsymbol{Y} = \boldsymbol{y})$$

Note that we need only k x n numbers ($\mathrm{P}(\boldsymbol{X}_i = \boldsymbol{x}_i \mid \boldsymbol{Y} = \boldsymbol{y})$) to implement this classifier, instead of k^n if we were to use the full model, without independence assumption.

Naïve Bayes Implementation

- Small (but important) implementation detail: If *n* is large, we may be taking the product of a large number of small floating-point values → underflow → avoided by taking the log.
- Take the max of:

$$\log \mathrm{P}(\boldsymbol{X}_1 = \boldsymbol{x}_1 \mid \boldsymbol{Y} = \boldsymbol{y}) + \ldots + \log \mathrm{P}(\boldsymbol{X}_n = \boldsymbol{x}_n \mid \boldsymbol{Y} = \boldsymbol{y}) + \log \mathrm{P}(\boldsymbol{Y} = \boldsymbol{y})$$

- Instead of:

$$\mathrm{P}(\boldsymbol{X}_1 = \boldsymbol{x}_1 \mid \boldsymbol{Y} = \boldsymbol{y}) \ldots \mathrm{P}(\boldsymbol{X}_n = \boldsymbol{x}_n \mid \boldsymbol{Y} = \boldsymbol{y}) \mathrm{P}(\boldsymbol{Y} = \boldsymbol{y})$$

Same Example, the Naïve Bayes Way

- Three variables:
  - Hair = {blond, dark}
  - Height = {tall,short}
  - Country = {Gromland, Polvia}
- Training data: Values of (Eye,Height,Country) collected over population

(B,T,G)	(B,T,G)	(B,T,P)
(D,T,G)	(D,T,G)	(B,T,P)
(D,T,G)	(D,T,G)	(B,T,P)
(D,T,G)	(D,T,G)	(D,T,P)
(B,T,G)	(B,T,G)	(D,T,P)
(B,S,G)	(B,S,G)	(D,S,P)
(B,S,G)	(B,S,G)	(B,S,P)
(D,S,G)	(D,S,G)	(D,S,P)

P(B,T|G)P(G) ≈ P(B|G)P(T|G)P(G)

8/16 x 10/16 x 2/3 ≈ 160/768 = 40/192

P(B,T|P)P(P) ≈ 4/8 x 5/8 x 1/3 = 20/192

Conclusion: Country = G

Same Exa

- Three vari
  - Hair = {bl
  - Height =
  - Country = {Gromland, Polvia}
- Training data: Values of (Eye Heig ry)

The values are of course different, but the conclusion remains the same
0.17 vs. 0.2 for Country = G
0.125 vs. 0.1 for Country = P

The variables are not independent so it is only an approximation.

(B,T,G)	(B,T,G)	(B,T,P)
(D,T,G)	(D,T,G)	(B,T,P)
(D,T,G)	(D,T,G)	(B,T,P)
(D,T,G)	(D,T,G)	(D,T,P)
(B,T,G)	(B,T,G)	(D,T,P)
(B,S,G)	(B,S,G)	(D,S,P)
(B,S,G)	(B,S,G)	(B,S,P)
(D,S,G)	(D,S,G)	(D,S,P)

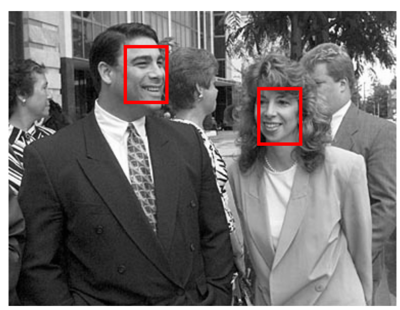
P(B,T| ≈ P(B|G) G)P(G)
8/16 x 10/ 2/3 ≈ 16 8 = 40/192

P(B,T|P)P(P) ≈ 4/8 x 5/ x 1/3 = 20/192

Conclusion: Country = G

Bayes at Work: Face Detection

Input Image

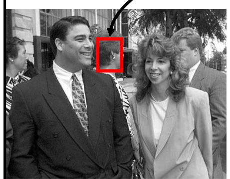


Find the faces (quickly)

Approach:
- Model the likelihood of an image window assuming face/non-face
- Use independence assumption along the way to make computations tractable

Source: Adapted from work by Henry Schneiderman (CMU & Pittsburgh Pattern Recognition)
http://vasc.ri.cmu.edu/cgi-bin/demos/findface.cgi

Naïve Bayes at Work



Move a window over an input image

At every position of the window:

1. Compute the values $x_1,..,x_n$ of a bunch of features $X_1,..,X_n$ from the image content within the window
2. Retrieve the probabilities:

$$P(X_i = x_i \mid \mathrm{Face}),\ P(X_i = x_i \mid \neg\mathrm{Face})\ i = 1,\ldots,n$$

from tables learned off-line

3. Assuming independence, compute:

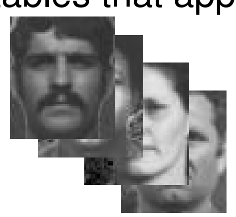
(1) $\mathrm{P}(\mathrm{Face})\,\mathrm{P}(X_1 = x_1 \mid \mathrm{Face})\ldots\mathrm{P}(X_n = x_n \mid \mathrm{Face})$

(2) $\mathrm{P}(\neg\mathrm{Face})\,\mathrm{P}(X_1 = x_1 \mid \neg\mathrm{Face})\ldots\mathrm{P}(X_n = x_n \mid \neg\mathrm{Face})$

4.Classify the window as a face if (1) > (2)

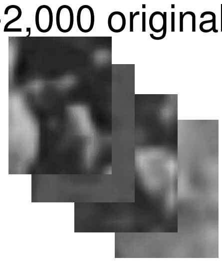
Learning

- Collect the values of the features for training data in tables that approximate the probabilities

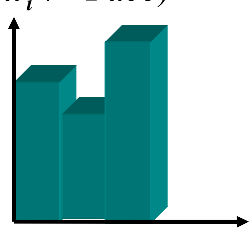


$P(X_i = x_i \mid \mathrm{Face})$

Face examples:
50-2,000 original images

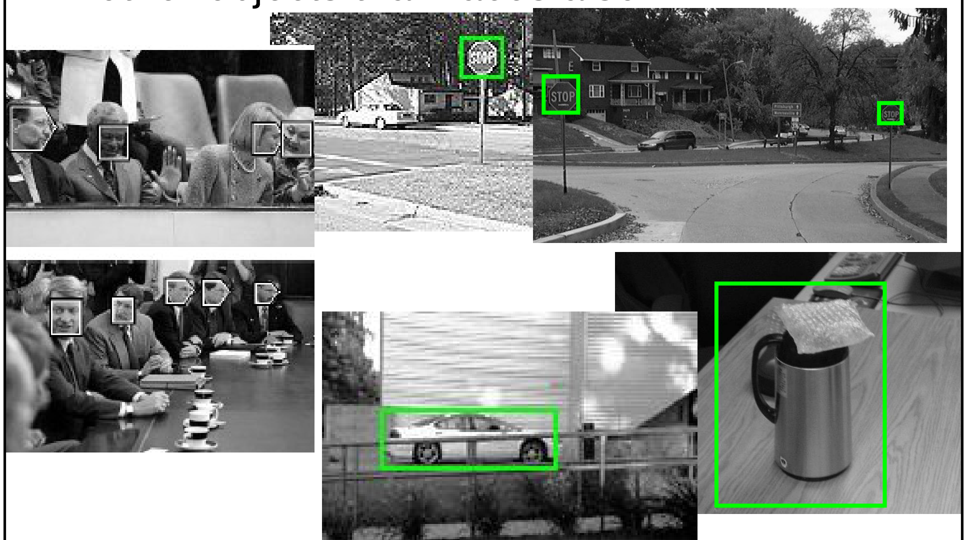


$P(X_i = x_i \mid \neg\mathrm{Face})$



Non-face examples:
~10,000,000 examples

- Yes it works. And in real-time. And with other objects than faces also….



Naïve

Move a

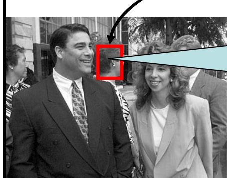
At every

1. Comp

of features $X_1,\ldots,X_n$ from the image content

within the window

2. Retrieve the probabilities:



The independence assumption is clearly violated: All the features X_i are computed from the same data (the image data inside the window). Nonetheless, the approximation is good enough for the classifier to work.

$$\boldsymbol{P}(\boldsymbol{X}_i = \boldsymbol{x}_i \mid \text{Face}), \boldsymbol{P}(\boldsymbol{X}_i = \boldsymbol{x}_i \mid \neg\text{Face})\ \boldsymbol{i} = 1,\ldots,\boldsymbol{n}$$

from tables learned off-line

g independence, compute:

$$\text{Face})\ldots \text{P}(\boldsymbol{X}_n = \boldsymbol{x}_n \mid \text{Face})$$

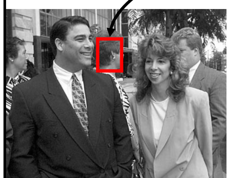
$$(\quad \neg\text{Face})\ldots \text{P}(\boldsymbol{X}_n = \boldsymbol{x}_n \mid \neg\text{Face})$$

he window as a face if (1) > (2)

If we did not make the independence assumption, we would have to learn K^n joint probabilities, where K is the number of possible values for each feature!

Naïve Bayes at Work

Move a window over an input image



At every position of the window:

1. Compute the values $x_1,..,x_n$ of a bunch of features $X_1,..,X_n$ from the image content within the window
2. Retrieve the probabilities:

$$P(X_i = x_i \mid \text{Face}),\, P(X_i = x_i \mid \neg\text{Face})\ i = 1,\ldots,n$$

from tables learned off-line

3. Assuming independence, compute:

(1) $\text{P}(\text{Face})\,\text{P}(X_1 = x_1 \mid \text{Face})\ldots\text{P}(X_n = x_n \mid \text{Face})$

(2) $\text{P}(\neg\text{Face})\,\text{P}(X_1 = x_1 \mid \neg\text{Face})\ldots\text{P}(X_n = x_n \mid \neg\text{Face})$

4.Classify the window as a face if (1) > (2)

Summary

- Basic probability concepts
- Bayes rule
- What are joint distributions
- Inference using joint distributions
- Learning joint distributions from data
- Independence
- Bayes classifiers
- Naïve Bayes approach