The Naïve Bayes Algorithm

Bayes Rule

Bayes Rule:

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

$P(A)$ prior

$P(A|B)$ posterior

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

…by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to a sure foundation for all our reasonings concerning past facts, and what is likely to be hereafter…. necessary to be considered by any that would give a clear account of the strength of *analogical* or *inductive reasoning…*

Applying Bayes Rule

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

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

A = you got flu   B = you just coughed

$P(A) = 0.05$, $P(B|A) = 0.8$, $P(B|\overline{A}) = 0.2$

What is P(flu|cough)=P(A|B)?

What does this has to do with function approximation?

Instead of learning $F: X \rightarrow Y$, learn $P(Y|X)$.

Can design algorithms that learn functions with uncertain outcomes (e.g., predicting tomorrow's stock price) **and that incorporate prior knowledge to guide learning** (e.g., a bias that tomorrow's stock price is likely to be similar to today's price).

The Joint Distribution

Example: Boolean variables A,B,C

- The key to building probabilistic models is to define a set of random variables, and to consider the joint probability distribution over them.

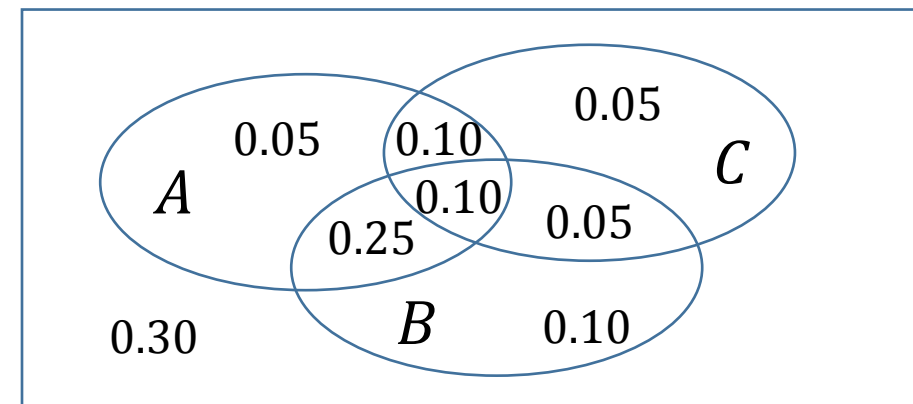| A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

The Joint Distribution

Example: Boolean variables A,B,C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values (M Boolean variables $\rightarrow 2^M$ rows).

2. For each combination of values, say how probable it is.

3. By the axioms of probability, these probabilities must sum to 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

Once we have the Joint Distribution, can ask for the probability of **any** logical expression involving these variables

College Degree	Hours worked	Wealth	prob
No	40.5-	Medium	0.253122
No	40.5-	Rich	0.0245895
No	40.5+	Medium	0.0421768
No	40.5+	Rich	0.0116293
Yes	40.5-	Medium	0.331313
Yes	40.5-	Rich	0.0971295
Yes	40.5+	Medium	0.134106
Yes	40.5+	Rich	0.105933

$$P(E) = \sum_{\text{rows matching E}} P(\text{row})$$

Using the Joint Distribution

Once we have the Joint Distribution, can ask for the probability of **any** logical expression involving these variables

P(College & Medium) = 0.4654

College Degree	Hours worked	Wealth	prob
No	40.5-	Medium	0.253122
No	40.5-	Rich	0.0245895
No	40.5+	Medium	0.0421768
No	40.5+	Rich	0.0116293
Yes	40.5-	Medium	0.331313
Yes	40.5-	Rich	0.0971295
Yes	40.5+	Medium	0.134106
Yes	40.5+	Rich	0.105933

$$P(E) = \sum_{\text{rows matching E}} P(\text{row})$$

Using the Joint Distribution

Once we have the Joint Distribution, can ask for the probability of **any** logical expression involving these variables

$$P(\text{Medium}) = 0.7604$$

College Degree	Hours worked	Wealth	prob
No	40.5-	Medium	0.253122
No	40.5-	Rich	0.0245895
No	40.5+	Medium	0.0421768
No	40.5+	Rich	0.0116293
Yes	40.5-	Medium	0.331313
Yes	40.5-	Rich	0.0971295
Yes	40.5+	Medium	0.134106
Yes	40.5+	Rich	0.105933

$$P(E) = \sum_{\text{rows matching E}} P(\text{row})$$

Inference with the Joint Distribution

Once we have the Joint Distribution, can ask for the probability of **any** logical expression involving these variables

College Degree	Hours worked	Wealth	prob
No	40.5-	Medium	0.253122
No	40.5-	Rich	0.0245895
No	40.5+	Medium	0.0421768
No	40.5+	Rich	0.0116293
Yes	40.5-	Medium	0.331313
Yes	40.5-	Rich	0.0971295
Yes	40.5+	Medium	0.134106
Yes	40.5+	Rich	0.105933

$$P(\text{College} \mid \text{Medium}) = \frac{0.4654}{0.7604} = 0.612$$

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

Learning and the Joint Distribution

Suppose we want to learn the function f: ⟨C, H⟩ → W

Equivalently, $P(W \mid C, H)$

One solution: learn joint distribution from data, calculate $P(W \mid C, H)$

College Degree	Hours worked	Wealth	prob
No	40.5-	Medium	0.253122
No	40.5-	Rich	0.0245895
No	40.5+	Medium	0.0421768
No	40.5+	Rich	0.0116293
Yes	40.5-	Medium	0.331313
Yes	40.5-	Rich	0.0971295
Yes	40.5+	Medium	0.134106
Yes	40.5+	Rich	0.105933

$$\text{e.g., } P(W = rich | C = no, H = 40.5 -) = \frac{0.0245895}{0.0245895 + 0.253122}$$

Idea: learn classifiers by learning P(Y | X)

Consider Y = Wealth

X = ⟨CollegeDegree, HoursWorked⟩

College Degree	Hours worked	Wealth	prob
No	40.5-	Medium	0.253122
No	40.5-	Rich	0.0245895
No	40.5+	Medium	0.0421768
No	40.5+	Rich	0.0116293
Yes	40.5-	Medium	0.331313
Yes	40.5-	Rich	0.0971295
Yes	40.5+	Medium	0.134106
Yes	40.5+	Rich	0.105933

College Degree	Hours worked	P(rich|C,HW)	P(medium|C,HW)
No	< 40.5	.09	.91
No	> 40.5	.21	.79
Yes	< 40.5	.23	.77
Yes	> 40.5	.38	.62

One approach: use this representation to learn P(Y|X).

Are we done?!?

One approach: use this representation to learn P(Y|X).

Main problem: learning P(Y|X) might require more data than we have…

Example:

Consider learning joint distributions with 100 attributes

Number of rows in this table? $2^{100} \sim 100^{10} \sim 10^{30}$

Number of people on Earth? 10^9

Fraction of rows with 0 training examples: 0.9999

What to do?

1. Be smart about how to estimate probabilities

2. Be smart about how to represent joint distributions

Be smart about how to estimate probabilities

Principle 1: Maximum Likelihood Estimation

Choose parameter $\hat{\theta}$ that maximizes likelihood of observed data $P(data|\hat{\theta})$

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_T + \alpha_H}$$

Can switch between these with Bayes Theorem

Principle 2: Maximum Aposteriori Probability

Choose parameter $\hat{\theta}$ that maximizes likelihood the posterior prob $P(\hat{\theta}|data)$

$$\hat{\theta}_{MAP} = \frac{\alpha_H + \#halucinated\_Hs}{(\alpha_T + \#halucinated\_Ts) + (\alpha_H + \#halucinated\_Hs)}$$

Be smart about how to represent joint distributions

Naïve Bayes algorithms assumes that

$$P(X_1, X_2, \ldots, X_n | Y) = \prod_i P(X_i | Y)$$

i.e., X_i and X_j are conditionally independent given Y, for all $i \neq j$

Conditional Independence

Definition

X is conditionally independent of Y given Z iff

    the probability distribution governing X is independent of Y, given the value of Z.

$$(\forall\, x, y, z): \quad P(X = x | Y = y, Z = z) = P(X = x | Z = z)$$

We often write as $P(X|Y, Z) = P(X|Z)$

E.g., $P(\text{Thunder}|\text{Rain}, \text{Lightening}) = P(\text{Thunder}|\text{Lightening})$

Note: does NOT **mean** that Thunder if independent of Rain.

Conditional Independence

X is conditionally independent of Y given Z iff

the probability distribution governing X is independent of Y, given the value of Z.

$$(\forall\, x, y, z): \quad P(X = x | Y = y, Z = z) = P(X = x | Z = z)$$

E.g., 3 Boolean random variables to describe the weather: Thunder, Rain, Lightening.

$$P(\text{Thunder} | \text{Rain}, \text{Lightening}) = P(\text{Thunder} | \text{Lightening})$$

Thunder is independent of Rain given Lightning. Lightning causes Thunder, once we know whether or not there is Lightning, no additional information about Thunder is provided by the value of Rain.

It does NOT mean that Thunder if independent of Rain.

Clear dependence of Thunder on Rain in general, but there is no conditional dependence once we know the value of Lightning.

Conditional Independence

Definition

X is **conditionally independent** of **Y** given **Z** iff

the probability distribution governing X is independent of Y, given the value of Z.

$$(\forall\, x, y, z): \quad P(X = x | Y = y, Z = z) = P(X = x | Z = z)$$

We often write as $P(X | Y, Z) = P(X | Z)$

Equivalent to $P(X, Y | Z) = P(X | Z)P(Y | Z)$

Conditional Independence

Claim

X is **conditionally independent** of **Y** given **Z** iff $\quad P(X|Y, Z) = P(X|Z)$

Equivalent to $\quad P(X, Y|Z) = P(X|Z)P(Y|Z)$

$$P(X, Y|Z) = P(X|Y, Z)P(Y|Z)$$

$$= P(X|Z)P(Y|Z)$$

Conditional Independence

Claim

If X_i and X_j are conditionally independent given Y, for all $i \neq j$

$$P(X_1, X_2, \ldots, X_n | Y) = \prod_i P(X_i | Y)$$

If $X_1, \ldots, X_n, Y$ are all Boolean, how many parameters do we need to describe $P(X_1, X_2, \ldots, X_n | Y)$ and $P(Y)$?

- Without the conditional independence assumption: $2(2^n - 1) + 1$

- With conditional independence assumption: $2n + 1$

Naïve Bayes in a Nutshell

Bayes Rule: $P(Y = y_k | X_1, ..., X_n) = \dfrac{P(Y = y_k) P(X_1, ..., X_n | Y = y_k)}{P(X)}$

If X_i and X_j are conditionally independent given Y, for all $i \neq j$

$$P(Y = y_k | X_1, ..., X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{P(X)}$$

So, to pick the most probably Y for $X^{new} = (X_1^{new}, X_2^{new}, ..., X_n^{new})$

$$Y^{new} = \text{argmax}_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

Naïve Bayes: discrete X_i

Training phase (input: training examples)

- For each value y_k, estimate $\pi_k = P(Y = y_k)$; get $\widehat{\pi_k}$

- For each value x_{ij} of attribute X_i estimate $\theta_{i,j,k} = P(X_i = x_{ij} | Y = y_k)$; get $\widehat{\theta_{i,j,k}}$

Testing phase:

- Classify $X^{new} = (X_1^{new}, X_2^{new}, \ldots, X_n^{new})$

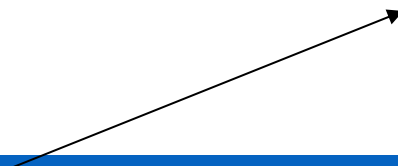$$Y^{new} = \text{argmax}_{y_k} \widehat{\pi_k} \prod_i \widehat{\theta_{i,new,k}}$$

[Ideal rule: $Y^{new} = \text{argmax}_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$]

Estimating parameters Y, X_i discrete

Maximum Likelihood Estimation

- For each value y_k, get $\widehat{\pi_k} = \widehat{P}(Y = y_k) = \frac{\#D(Y=y_k)}{|D|}$

- For each value x_{ij} of attribute X_i estimate $\theta_{i,j,k} = P(X_i = x_{ij} | Y = y_k)$;

$$\text{get } \widehat{\theta_{i,j,k}} = \widehat{P}(X_i = x_{ij} | Y = y_k) = \frac{\#D\big(X_i = x_{ij} \wedge Y = y_k\big)}{\#D(Y=y_k)}$$

Number of items in dataset D for which $Y = y_k$

Sublety 1: Violation of the Naïve Bayes Assumption

- Usually features are not conditionally independent given the label

$$P(X_1, X_2, \ldots, X_n | Y) \neq \prod_i P(X_i | Y)$$

- Nonetheless, NB is widely used:

  - NB often performs well, even when assumption is violated
  - [Domingos & Pazzani '96] discuss some conditions for good performance

Subtlety 2: Insufficient Training Data

- What if we never see a training instance where $X_1 = a$ and $Y = b$?

$$\text{e.g., } Y = \text{SpamEmail}, X = \text{"Earn"}$$

$$\widehat{P}(X_1 = a | Y = b) = 0 \qquad \widehat{P}(X_1 = a | Y = b) = 0$$

- Thus no matter what the values $X_2, \ldots, X_n$ take, we obtain:

$$\widehat{P}(Y = b | X_1 = a, X_2, \ldots, X_n) = 0$$

$$\widehat{P}(X_1 = a, X_2, \ldots, X_n | Y) = \widehat{P}(X_1 = a) \prod_{i \neq 1} P(X_i | Y)$$

- Solution: use MAP estimate!!!!

Estimating parameters Y, X_i discrete

Maximum Likelihood Estimation

- For each value y_k, get $\widehat{\pi_k} = \widehat{P}(Y = y_k) = \dfrac{\#D(Y=y_k)+l}{|D|+lK}$

- For each value x_{ij} of attribute X_i estimate $\theta_{i,j,k} = P(X_i = x_{ij}|Y = y_k)$;

$$\text{get } \widehat{\theta_{i,j,k}} = \widehat{P}(X_i = x_{ij}|Y = y_k) = \frac{\#D\left(X_i=x_{ij}\wedge Y=y_k\right)+l}{\#D(Y=y_k)+lJ}$$

J - number of distinct values that feature i can take; l determines the strength of this smoothing; assume the hallucinated examples are spread evenly over the possible values of X_i; so, number of hallucinated examples is lJ.

What you should know…

- Naïve Bayes classifier
  - What's the assumption
  - Why we use it
  - How do we learn it
  - Why is Bayesian estimation important