SNS COLLEGE OF TECHNOLOGY

Coimbatore – 35

An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A++' Grade

Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

Introduction to Bayesian Learning

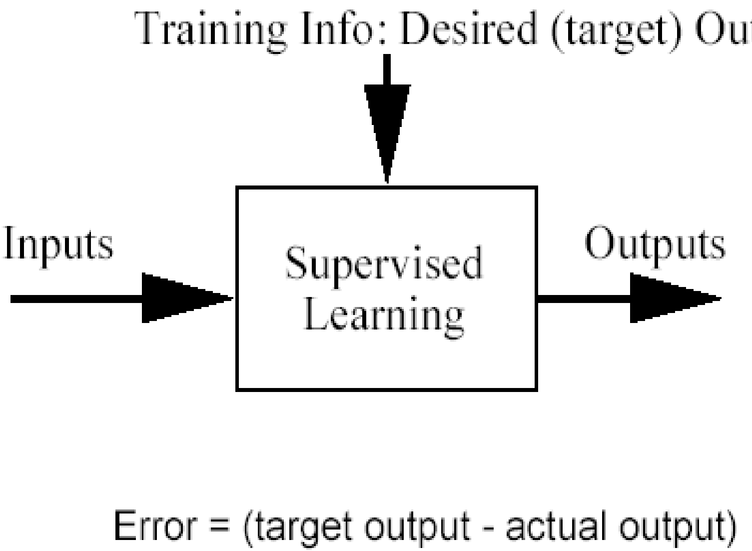
Overview

Today we learn about:

- Bayes rule & turn this into a classifier
  - E.g. How to decide if a patient is ill or healthy, based on
    - A probabilistic model of the observed data
    - Prior knowledge

Classification problem

- Training data: examples of the form (d,h(d))
  - where d are the data objects to classify (inputs)
  - and h(d) are the correct class info for d, $h(d) \in \{1,\ldots K\}$
- Goal: given d_{new}, provide $h(d_{new})$



A word about the Bayesian framework

- Allows us to combine observed data and prior knowledge
- Provides practical learning algorithms
- It is a generative (model based) approach, which offers a useful conceptual framework
  - This means that any kind of objects (e.g. time series, trees, etc.) can be classified, based on a probabilistic model specification

Bayes' Rule

$$p(h \mid d) = \frac{P(d \mid h)P(h)}{P(d)}$$

Understanding Bayes' rule

d = data

h = hypothesis

Proof. Just rearrange :

$$p(h \mid d)P(d) = P(d \mid h)P(h)$$

$$P(d,h) = P(d,h)$$

the same joint probability on both sides

<u>Who is who in Bayes' rule</u>

$P(h)$: prior

$P(d \mid h)$: likelihood

$P(d) = \sum_h P(d \mid h)P(h)$: evidence

$p(h \mid d)$: posterior

Probabilities – auxiliary slide for memory refreshing

- Have two dice h_1 and h_2
- The probability of rolling an *i* given die h_1 is denoted $P(i|h_1)$. This is a ***conditional probability***
- Pick a die at random with probability $P(h_j)$, j=1 or 2. The probability for picking die h_j *and* rolling an i with it is called ***joint probability*** and is $P(i, h_j)=P(h_j)P(i|h_j)$.
- For any events X and Y, P(X,Y)=P(X|Y)P(Y)
- If we know P(X,Y), then the so-called ***marginal probability*** P(X) can be computed as $P(X)=\sum_Y P(X,Y)$
- Probabilities sum to 1. Conditional probabilities sum to 1 **provided that their conditions are the same**.

Does patient have cancer or not?

- A patient takes a lab test and the result comes back positive. It is known that the test returns a correct positive result in only 98% of the cases and a correct negative result in only 97% of the cases. Furthermore, only 0.008 of the entire population has this disease.

  1. What is the probability that this patient has cancer?
  2. What is the probability that he does not have cancer?
  3. What is the diagnosis?

$$\left.\begin{array}{l} hypothesis1: 'cancer' \\ hypothesis2: '\neg cancer' \end{array}\right\} \quad hypothesis\ space\ H$$

$$-data: '+'$$

$$1.\ P(cancer|+) = \frac{P(+|cancer)P(cancer)}{P(+)} = \frac{.....................}{.....................} = ...$$

$$P(+|cancer) = 0.98$$

$$P(cancer) = 0.008$$

$$P(+) = P(+|cancer)P(cancer) + P(+|\neg cancer)P(\neg cancer)$$

$$= ....$$

$$P(+|\neg cancer) = 0.03$$

$$P(\neg cancer) = ...$$

$$2.\ P(\neg cancer|+) = ...$$

$$3.\ Diagnosis??$$

Choosing Hypotheses

- *Maximum Likelihood* hypothesis:

$$h_{ML} = \arg\max_{h \in H} P(d|h)$$

- Generally we want the most probable hypothesis given training data.This is the *maximum a posteriori* hypothesis:
  - Useful observation: it does not depend on the denominator P(d)

$$h_{MAP} = \arg\max_{h \in H} P(h|d)$$

Now we compute the diagnosis

– To find the Maximum Likelihood hypothesis, we evaluate P(d|h) for the data d, which is the positive lab test and chose the hypothesis (diagnosis) that maximises it:

$$P(+|cancer) = \ldots$$

$$P(+|\neg cancer) = \ldots$$

$$\Rightarrow Diagnosis_{ML} = \ldots$$

– To find the Maximum A Posteriori hypothesis, we evaluate P(d|h)P(h) for the data d, which is the positive lab test and chose the hypothesis (diagnosis) that maximises it. This is the same as choosing the hypotheses gives the higher posterior probability.

$$P(+|cancer)P(cancer) = \ldots$$

$$P(+|\neg cancer)P(\neg cancer) = \ldots$$

$$\Rightarrow Diagnosis_{MAP} = \ldots$$

The Naïve Bayes Classifier

- What can we do if our data *d* has several attributes?
- Naïve Bayes assumption: Attributes that describe data instances are conditionally independent given the classification hypothesis

$$P(\mathbf{d}|h) = P(a_1, \ldots, a_T|h) = \prod_t P(a_t|h)$$

  - it is a simplifying assumption, obviously it may be violated in reality
  - in spite of that, it works well in practice
- The Bayesian classifier that uses the Naïve Bayes assumption and computes the MAP hypothesis is called Naïve Bayes classifier
- One of the most practical learning methods
- Successful applications:
  - Medical Diagnosis
  - Text classification

Example. ‘Play Tennis’ data

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Naïve Bayes solution

Classify any new datum instance $\mathbf{x}=(a_1,\ldots a_T)$ *as:*

$$h_{Naive\ Bayes} = \arg\max_h P(h) \prod_t P(a_t|h)$$

- To do this based on training examples, we need to estimate the parameters from the training examples:
  - For each target value (hypothesis) *h*

$$\hat{P}(h) := \text{estimate } P(h)$$

$$\hat{P}(a_t|h) := \text{estimate } P(a_t|h)$$

  - For each attribute value a_t of each datum instance

Based on the examples in the table, classify the following datum ***x***:

x=(Outl=Sunny, Temp=Cool, Hum=High, Wind=strong)

- That means: Play tennis or not?

$$h_{NB} = \arg\max_{h \in [yes, no]} P(h)P(\mathbf{x}|h) = \arg\max_{h \in [yes, no]} P(h)\prod_t P(a_t|h)$$

$$= \arg\max_{h \in [yes, no]} P(h)P(Outlook = sunny|h)P(Temp = cool|h)P(Humidity = high|h)P(Wind = strong|h)$$

- Working:

$$P(PlayTennis = yes) = 9/14 = 0.64$$

$$P(PlayTennis = no) = 5/14 = 0.36$$

$$P(Wind = strong|PlayTennis = yes) = 3/9 = 0.33$$

$$P(Wind = strong|PlayTennis = no) = 3/5 = 0.60$$

$$etc.$$

$$P(yes)P(sunny|yes)P(cool|yes)P(high|yes)P(strong|yes) = 0.0053$$

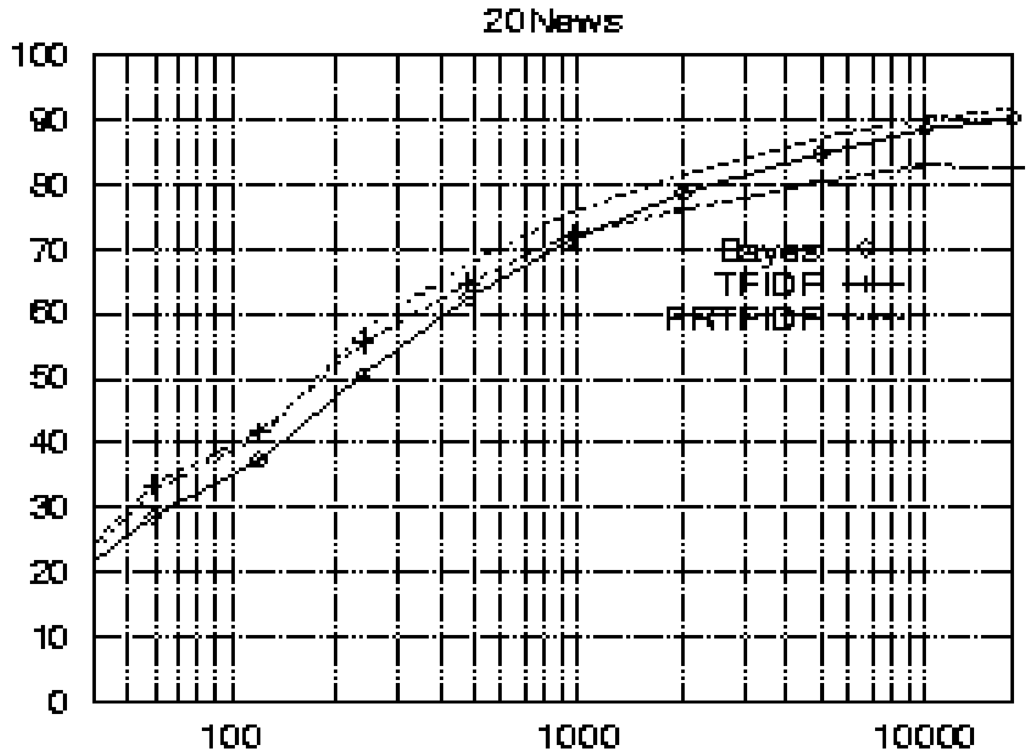
$$P(no)P(sunny|no)P(cool|no)P(high|no)P(strong|no) = 0.0206$$

$$\Rightarrow answer: PlayTennis(x) = no$$

Learning to classify text

- Learn from examples which articles are of interest
- The attributes are the words
- Observe the Naïve Bayes assumption just means that we have a random sequence model within each class!
- NB classifiers are one of the most effective for this task
- Resources for those interested:
  - Tom Mitchell: Machine Learning (book) Chapter 6.

Results on a benchmark text corpus



Accuracy vs. Training set size (1/3 withheld for test)

Remember

- Bayes' rule can be turned into a classifier
- Maximum A Posteriori (MAP) hypothesis estimation incorporates prior knowledge; Max Likelihood doesn't
- Naive Bayes Classifier is a simple but effective Bayesian classifier for vector data (i.e. data with several attributes) that assumes that attributes are independent given the class.
- Bayesian classification is a generative approach to classification

Resources

- Textbook reading (contains details about using Naïve Bayes for text classification):

  Tom Mitchell, Machine Learning (book), Chapter 6.

- Further reading for those interested to learn more:

  http://www-2.cs.cmu.edu/~tom/NewChapters.html