COMPSCI762: Foundations of Machine Learning

Introduction to Bayesian Learning

Jörg Simon Wicker and Katerina Taškova

The University of Auckland

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

SCIENCE
SCHOOL OF COMPUTER SCIENCE

Bayesian Learning

Motivation and Introduction

Bayes' Theorem

Maximum A Posteriori Hypothesis

Bayes Optimal Classifier

Maximum Likelihood and Least-Squared Error

Minimum Description Length

Naive Bayes Classifier

  Naive Bayes for Document Classification

Bayesian Networks

Partly based on Mitchel's book, lecture slides from Stanford's NLP lecture and The University of Utah

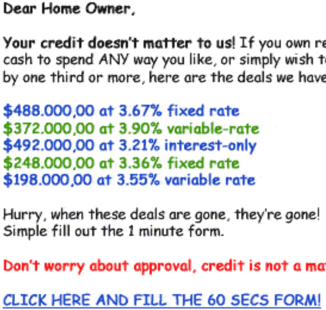
Motivation and Introduction

Spam Filtering

- Spam filters analyze the text of emails and classify them into Spam and Ham
  - SpamAssassin Features:
    - Mentions Generic Viagra
    - Online Pharmacy
    - Mentions millions of (dollar) ((dollar) NN,NNN,NNN.NN)
    - Phrase: impress ... girl
    - From: starts with many numbers
    - Subject is all capitals
    - HTML has a low ratio of text to image area
    - One hundred percent guaranteed
    - Claims you can be removed from the list
    - 'Prestigious Non-Accredited Universities'

Spam Filtering

- We want to build a system that detects e-mails
- Can we formulate this as a supervised learning task?



Dear Home Owner,

Your credit doesn't matter to us! If you own real estate and want IMMEDIATE cash to spend ANY way you like, or simply wish to LOWER your monthly payments by one third or more, here are the deals we have today:

$488.000,00 at 3.67% fixed rate
$372.000,00 at 3.90% variable-rate
$492.000,00 at 3.21% interest-only
$248.000,00 at 3.36% fixed rate
$198.000,00 at 3.55% variable rate

Hurry, when these deals are gone, they're gone!
Simple fill out the 1 minute form.

Don't worry about approval, credit is not a matter!

CLICK HERE AND FILL THE 60 SECS FORM!

Spam Filtering as Supervised Learning

- Collect a large number of e-mails, get users to label them

$	Hi	CS	762	Vicodin	Offer	...	Spam?
1	1	0	0	1	0	...	1
0	0	0	0	1	1	...	1
0	1	1	1	0	0	...	0
...	...	...	...	...	...	...	...

- We can use $(y_i = 1)$ if e-mail i is spam, $(y_i = 0)$ if e-mail is not spam
- Extract features of each e-mail (*bag of words*)
  - $(x_{ij} = 1)$ if word/phrase j is in e-mail i, $(x_{ij} = 0)$ if it is not

Feature Representation for Spam Filtering

- Are there better features than bag of words?
  - We can add bigrams (sets of two words)
    - “CS 762”, “Computer Science”
  - Or trigrams (sets of three words)
    - “University of Auckland”, “Limited time offer”
  - We might include the sender domain
    - <sender domain == “mail.com”>
  - We might include regular expressions

Probabilistic Classifiers

- For years, best spam filtering methods used naive Bayes
  - A probabilistic classifier based on Bayes rule
  - It tends to work well with bag of words
- Probabilistic classifiers model the conditional probability, $p(y_i|x_i)$
  - “If a message has word x_i, what is the probability that message is spam?”
- Classify it as spam if conditional probability of spam is higher than that of not spam
  - If $p(y_i = 1|x_i) > p(y_i = 0|x_i)$ return “spam”else return “not spam”

In the Church of the Reverend Bayes

- So far, learning as a search or based on rules
- For Bayesians: Learning is just another application of the Bayes' Theorem

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- Search the most likely hypothesis

Basic assumption

- Quantities of interest are governed by probability distributions
- Optimal decisions can be made by reasoning about these probabilities together with observed training data

Relevance

- Bayesian Learning is relevant for two reasons
  1. Explicit manipulation of probabilities
     - Among the most practical approaches to certain types of learning problems
     - E.g. Bayes classifier is competitive with decision tree and ANNs
  2. Useful framework for understanding learning methods that do not explicitly manipulate probabilities
     - Determine conditions under which algorithms output the most probable hypothesis
     - E.g. justification of the least square error function
     - E.g. justification of the why smaller decision trees are preferred (Occam's razor)

Practical difficulties

- Initial knowledge of many probabilities is required
- Significant computational costs required

Bayes' Theorem

Bayes' Theorem

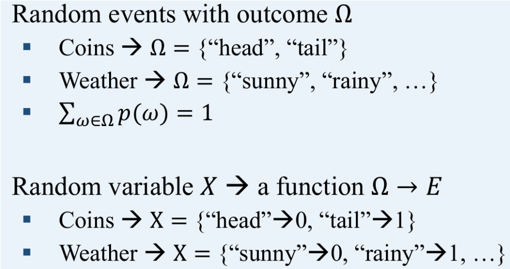
- Machine Learning is interested in the best hypothesis h from some space H, given observed training data D
- best hypothesis $\approx$ most probable hypothesis
- Bayes' Theorem provides a direct method of calculating the probability of such a hypothesis based on
  - its prior probability,
  - the probabilities of observing various data given the hypothesis, and
  - the observed data itself

Bayes' Theorem

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

which is short for $\forall x, y$:

$$P(Y = y|X = x) = \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)}$$



Random events with outcome Ω

- Coins → $\Omega = \{$“head”, “tail”$\}$
- Weather → $\Omega = \{$“sunny”, “rainy”, …$\}$
- $\sum_{\omega \in \Omega} p(\omega) = 1$

Random variable X → a function $\Omega \to E$

- Coins → X = {“head”→0, “tail”→1}
- Weather → X = {“sunny”→0, “rainy”→1, …}

Probability: random events

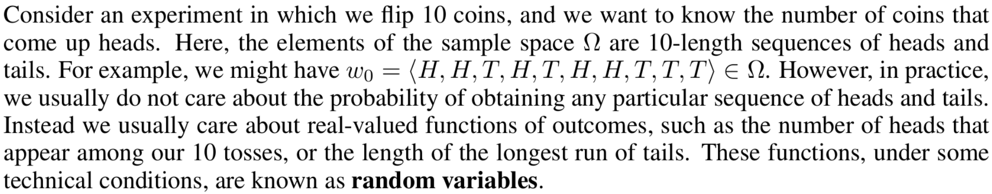
- **Sample space** Ω: The set of all the outcomes of a random experiment. Here, each outcome $\omega \in \Omega$ can be thought of as a complete description of the state of the real world at the end of the experiment.
- **Set of events** (or **event space**) $\mathcal{F}$: A set whose elements $A \in \mathcal{F}$ (called **events**) are subsets of Ω (i.e., $A \subseteq \Omega$ is a collection of possible outcomes of an experiment).[1].
- **Probability measure**: A function $P : \mathcal{F} \rightarrow \mathbb{R}$ that satisfies the following properties,
  - $P(A) \geq 0$, for all $A \in \mathcal{F}$
  - $P(\Omega) = 1$
  - If $A_1, A_2, \ldots$ are disjoint events (i.e., $A_i \cap A_j = \emptyset$ whenever $i \neq j$), then

$$P(\cup_i A_i) = \sum_i P(A_i)$$

Probability: random events

Example: Consider the event of tossing a six-sided die. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. We can define different event spaces on this sample space. For example, the simplest event space is the trivial event space $\mathcal{F} = \{\emptyset, \Omega\}$. Another event space is the set of all subsets of Ω. For the first event space, the unique probability measure satisfying the requirements above is given by $P(\emptyset) = 0, P(\Omega) = 1$. For the second event space, one valid probability measure is to assign the probability of each set in the event space to be $\frac{i}{6}$ where i is the number of elements of that set; for example, $P(\{1, 2, 3, 4\}) = \frac{4}{6}$ and $P(\{1, 2, 3\}) = \frac{3}{6}$.

Probability: random variables



Consider an experiment in which we flip 10 coins, and we want to know the number of coins that come up heads. Here, the elements of the sample space Ω are 10-length sequences of heads and tails. For example, we might have $w_0 = \langle H, H, T, H, T, H, H, T, T, T \rangle \in \Omega$. However, in practice, we usually do not care about the probability of obtaining any particular sequence of heads and tails. Instead we usually care about real-valued functions of outcomes, such as the number of heads that appear among our 10 tosses, or the length of the longest run of tails. These functions, under some technical conditions, are known as **random variables**.

Bayesian Learning

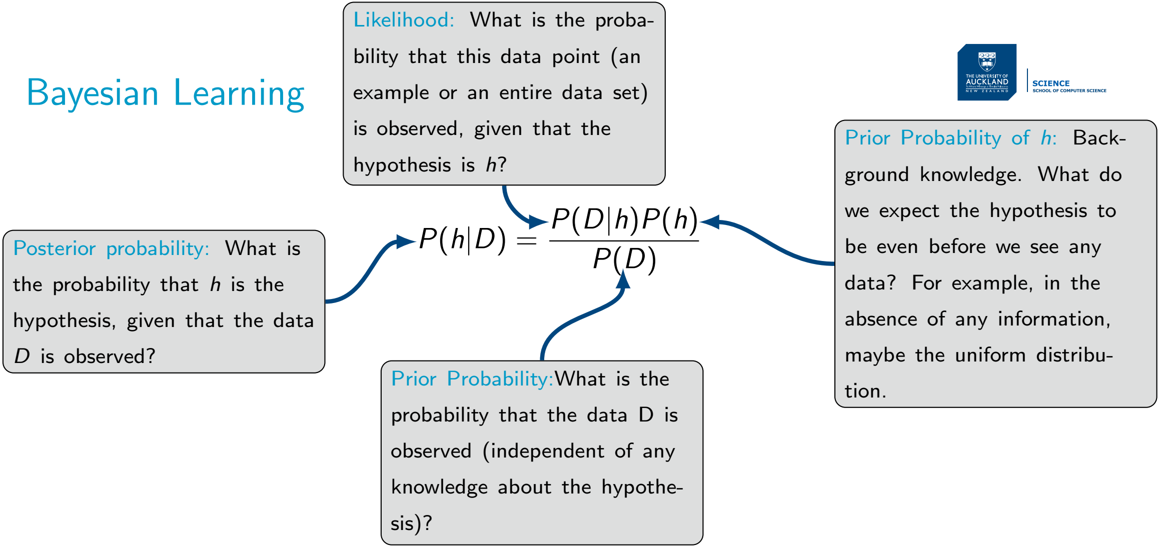
- Given data set D, we want to find the *best* hypothesis h
- What does "*best*" mean?
- Bayesian learning uses $P(h|D)$, the conditional probability of a hypothesis given the data, to define *best*

Bayesian Learning

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- Key insights: Both h and D are events
  - D: The event that we observed *this* particular data set
  - h: The event that the hypothesis h is the true hypothesis
- $\Rightarrow$ Hypothesis h and observed data set D to substitute X and Y in Bayes' Theorem

Bayesian Learning



$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Posterior probability: What is the probability that h is the hypothesis, given that the data D is observed?

Likelihood: What is the probability that this data point (an example or an entire data set) is observed, given that the hypothesis is h?

Prior Probability of h: Background knowledge. What do we expect the hypothesis to be even before we see any data? For example, in the absence of any information, maybe the uniform distribution.

Prior Probability:What is the probability that the data D is observed (independent of any knowledge about the hypothesis)?

Maximum A Posteriori Hypothesis

Maximum A Posteriori Hypothesis

- In many learning scenarios, the learner considers some set of candidate hypotheses H and is interested in finding the most probable hypothesis $h \in H$ given the observed training data D
- Any maximally probable hypothesis is called maximum a posteriori (MAP) hypothesis h_{MAP}

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(h|D) \\ &= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \text{by Bayes' Theorem} \\ &= \arg\max_{h \in H} P(D|h)P(h) \end{aligned}$$

- Note that we can drop $P(D)$ because it is a constant independent of h

Choosing a Hypothesis

- Sometimes it is assumed that every hypothesis is equally probable apriori
- In this case, the equation above can be simplified
- Because $P(D|H)$ is often called the likelihood of D given h, any hypothesis that maximizes $P(D|h)$ is called maximum likelihood (ML) hypothesis

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

- Note that in this case P(h) can be dropped, because it is constant for each $h \in H$

Example

- Consider a medical diagnosis problem in which there are two alternative hypotheses
  - The patient has a particular form of cancer (denoted by $cancer$)
  - The patient does not have a cancer (denoted by $\neg cancer$)
- The available data is from a particular laboratory with two possible outcomes: $\oplus$ (positive) and $\ominus$(negative)

$$\begin{aligned} P(cancer) &= 0.008 & P(\neg cancer) &= 0.992 \\ P(\oplus|cancer) &= 0.98 & P(\ominus|cancer) &= 0.02 \\ P(\oplus|\neg cancer) &= 0.03 & P(\ominus|\neg cancer) &= 0.97 \end{aligned}$$

- Suppose a new patient is observed for whom the lab test returns a positive ($\oplus$) result. Should we diagnose the patient as having cancer or not?

$$\begin{aligned} P(cancer|\oplus) &\sim P(\oplus|cancer)P(cancer) && = 0.98 * 0.008 = 0.0078 \\ P(\neg cancer|\oplus) &\sim P(\oplus|\neg cancer)P(\neg cancer) && = 0.03 * 0.992 = 0.0298 \\ \Rightarrow h_{MAP} &= \neg cancer \end{aligned}$$

Does the probability of cancer increase?

Example (cont.)

- The exact posterior probabilities can be determined by ?normalizing the above probabilities to sum up to 1

$$P(cancer|\oplus) = \frac{0.0078}{0.0078 + 0.0298} = 0.21$$
$$P(\neg cancer|\oplus) = \frac{0.0298}{0.0078 + 0.0298} = 0.79$$

The probability of cancer increases when we observe a positive test result!

Bayes Optimal Classifier

Bayes Optimal Classifier

- Question: What is the most probable classification of the new instance given the training data?
- Simply applying h_{MAP} is not the best solution (as one could wrongly think of)
- Example
  - $H = \{h_1, h_2, h_3\}$, where $P(h_1|D) = 0.4$, $P(h_2|D) = P(h_3|D) = 0.3$
  - $h_{MAP} = h_1$
  - Consider a new instance x encountered, which is classified positive by h_1 and negative by h_2 , h_3
  - Taking all hypotheses into account
    - The probability that x is positive is 0.4
    - The probability that x is negative is 0.6

$\Rightarrow$ most probable classification $\neq$ classification of h_{MAP}

Bayes Optimal Classifier

- The most probable classification is obtained by combining the predictions of all hypotheses, weighted by their posterior probabilities

$$P(v_j|D) = \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

  where $P(v_j|D)$ is the probability that the correct classification is v_j

- Bayes optimal classifier

$$\arg\max_{v_j \in V} P(v_j|D) = \arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

Example

$V = \{\oplus, \ominus\}$

$$P(h_1|D) = 0.4, P(\ominus|h_1) = 0, P(\oplus|h_1) = 1$$
$$P(h_2|D) = 0.3, P(\ominus|h_2) = 1, P(\oplus|h_2) = 0$$
$$P(h_3|D) = 0.3, P(\ominus|h_3) = 1, P(\oplus|h_3) = 0$$

therefore

$$P(\oplus|D) = \sum_{h_i \in H} P(\oplus|h_i)P(h_i|D) = 0.4$$
$$P(\ominus|D) = \sum_{h_i \in H} P(\ominus|h_i)P(h_i|D) = 0.6$$

and

$$\underset{v_j \in \{\oplus, \ominus\}}{\arg\max} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) = \ominus$$

Literature

- Chapter 6 of Mitchell's *Machine Learning*
- Chapter 8 of Bishop's *Pattern Recognition and Machine Learning*

Thank you for your attention!

https://ml.auckland.ac.nz