

# Bayesian Learning

Kaiqi Zhao  
The University of Auckland

*Slides are partially based on the materials from Mitchel's book and Stanford's NLP lectures*

# Content

- Motivation and Introduction
- Bayes' Theorem
- Maximum A Posteriori Hypothesis
- Maximum Likelihood and Least-Squared Error
- Minimum Description Length
- Bayes Optimal Classifier
- Naive Bayes Classifier
  - Naive Bayes for Document Classification
- Bayesian Networks

# 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 spam 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 words  $x_i$ , 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 objective function
    - E.g. justification of 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  $D$
- The best hypothesis  $\approx$  most probable hypothesis
- Bayes' Theorem provides a direct method of calculating the probability of such a hypothesis based on
  - The prior probability, and
  - The probabilities of observing various data given the hypothesis, and
  - The observed data itself

# Bayes' Theorem

- Two random events (or random variables)  $X$  and  $Y$ :

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

- 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  $\Omega$ 
  - Coins  $\rightarrow \Omega = \{\text{"head"}, \text{"tail"}\}$
  - Weather  $\rightarrow \Omega = \{\text{"sunny"}, \text{"rainy"}, \dots\}$
  - $\sum_{\omega \in \Omega} p(\omega) = 1$
- Random variable  $X \rightarrow$  a function  $\Omega \rightarrow E$ 
  - Coins  $\rightarrow X = \{\text{"head"} \rightarrow 0, \text{"tail"} \rightarrow 1\}$
  - Weather  $\rightarrow X = \{\text{"sunny"} \rightarrow 0, \text{"rainy"} \rightarrow 1, \dots\}$

# 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

- Use hypothesis  $h$  and observed data set  $D$  to substitute  $X$  and  $Y$  in Bayes' Theorem

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

- 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

# Bayes' Theorem

**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  $h$ ?

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

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

**Prior Probability of  $h$ :**  
Back-ground 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.

# 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  $h_{MAP}$

$$\begin{aligned}
 h_{MAP} &= \operatorname{argmax}_{h \in H} P(h|D) \\
 &= \operatorname{argmax}_{h \in H} \frac{P(D|h)P(h)}{P(D)} \quad - \text{By Bayes' Theorem} \\
 &= \operatorname{argmax}_{h \in H} P(D|h)P(h)
 \end{aligned}$$

- 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

$$h_{ML} = \operatorname{argmax}_{h \in H} P(D|h)$$

- 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)

$$P(cancer) = 0.008$$

$$P(\neg cancer) = 0.992$$

$$P(\oplus | cancer) = 0.98$$

$$P(\ominus | cancer) = 0.02$$

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

$$P(\ominus | \neg cancer) = 0.97$$

- 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?

$$P(cancer | \oplus) \propto P(\oplus | cancer)P(cancer) = 0.98 * 0.008 = 0.0078$$

$$P(\neg cancer | \oplus) \propto P(\oplus | \neg cancer)P(\neg cancer) = 0.03 * 0.992 = 0.0298$$

$$\Rightarrow h_{MAP} = \neg cancer$$

Does the probability of cancer increase?

# Example

- The exact posterior probabilities can be determined by normalizing the above probabilities to 1.

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

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

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