Lecture 2: Bayesian Inference

- Random variables and probabilities
- Beliefs
- Conditional probability and Bayes rule
- Independence of random variables
- Using Bayes rule for inference
- Conditional independence

Random variables and probability

- A **random variable** X describes an outcome that cannot be determined in advance (e.g. the roll of a die)
- The **sample space** S of a random variable X is the set of all possible values of the variable

  E.g. For a die, $S = \{1, 2, 3, 4, 5, 6\}$
- An **event** is a subset of S. E.g. $e = \{1\}$ corresponds to a die roll of 1
- Usually, random variables are still governed by some law of nature, described as a **probability function** p defined on S. $p(x)$ defines the chance that variable X takes value $x \in S$.

  E.g. for a die roll with a fair die, $p(1) = p(2) = \ldots = p(6) = \frac{1}{6}$

  Note: We still cannot determine the value of X, just the chance of encountering a given value

Discrete random variables

If X is a discrete variable, then a probability space $p(x)$ has the following properties:

$$0 \leq p(x) \leq 1, \forall x \in S \text{ and } \sum_{x \in S} p(x) = 1$$

Continuous random variables

- If X is a continuous random variable, its probability density function $p(x)$ has the following properties:

  $$0 \leq p(x), \forall x \in S \text{ and } \int p(x)dx = 1$$

  Note that in this case $p(x)$ can be greater than 1, because it is **not** a probability value

- For continuous variables, we can also define a **cumulative distribution function**, c, which takes values between 0 and 1:

  $$c(a) = \int_{-\infty}^{a} p(x)dx$$

  $c(a)$ is the probability that random variable X has value less than or equal to a.

Terminology

- The **n-th moment of a random variable** X is defined as:

$$M_n = \sum_{x \in S} x^n p(x)$$

- The first moment is called the **expectation** or **mean**:

$$E\{x\} = M_1 = \sum_{x \in S} x p(x)$$

E.g. for a roll with a fair die, the expectation is:

$$M_1 = \sum_{x \in \{1,2,3,4,5,6\}} x \frac{1}{6} = 3.5$$

Note: As illustrated above, the expectation is **not** the value we expect to see the most.

And more terminology...

- The **variance** is defined as:

$$Var\{x\} = M_2 - M_1^2 = E\{x^2\} - E\{x\}^2$$

- The **standard deviation** $\sigma = \sqrt{Var\{x\}}$ evaluates the "spread" of x with respect to its mean

Beliefs

- We will use probability in order to describe the world and the existing uncertainties
- **Beliefs** (also called Bayesian or subjective probabilities) relate logical propositions to the current state of knowledge
- Beliefs are **subjective** assertions about the world, given one's state of knowledge

  E.g. $P(\text{Some day AI agents will rule the world}) = 0.1$ reflects a personal belief, based on one's state of knowledge about current AI, technology trends, etc.
- Different agents may hold different beliefs
- **Prior (unconditional) beliefs** denote belief prior to the arrival of any new evidence.

Axioms of probability

Beliefs satisfy the axioms of probability.

For any propositions A, B:

1. $0 \leq P(A) \leq 1$
2. $P(True) = 1$
3. $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$, or equivalently,
   $P(A \vee B) = P(A) + P(B)$ if A and B are mutually exclusive

The axioms of probability limit the class of functions that can be considered probability functions.

Using functions that disobey these laws as probabilities can force suboptimal decisions (de Finetti, 1931).

Defining probabilistic models

- We define the world as a set of random variables $\Omega = \{X_1 \ldots X_n\}$.
- A **probabilistic model** is an encoding of probabilistic information that allows us to compute the probability of any event in the world

A simple probabilistic model:

- We divide the world into a set of elementary, mutually events, called **states**

  E.g. If the world is described by two Boolean variables A, B, a state will be a complete assignment of truth values for A and B.
- A **joint probability distribution function** assigns non-negative weights to each event, such that these weights sum to 1.

Inference using joint distributions

E.g. Suppose $Toothache$ and $Cavity$ are the random variables:

	$Toothache = true$	$Toothache = false$
$Cavity = true$	0.04	0.06
$Cavity = false$	0.01	0.89

The unconditional probability of any proposition is computable as the sum of entries from the full joint distribution

E.g. $P(Cavity) =$

$$P(Cavity, Toothache) + P(Cavity, \neg Toothache) = 0.1$$

Conditional probability

The basic statements in the Bayesian framework talk about **conditional probabilities**. $P(A|B)$ is the belief in event A given that event B is known with absolute certainty:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ if } P(B) \neq 0$$

Note that we can use either the set intersection or the logical “and” notation above.

The **product rule** gives an alternative formulation:

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

Bayes rule

Bayes rule is another alternative formulation of the product rule:

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

The **complete probability formula** states that:

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

or more generally,

$$P(A) = \sum_i P(A|b_i)P(b_i),$$

where b_i form a set of exhaustive and mutually exclusive events.

Chain rule

Chain rule is derived by successive application of product rule:

$$
\begin{aligned}
& P(X_1, \ldots, X_n) = \\
= \quad & P(X_1, \ldots, X_{n-1}) P(X_n | X_1, \ldots, X_{n-1}) \\
= \quad & P(X_1, \ldots, X_{n-2}) P(X_{n-1} | X_1, \ldots, X_{n-2}) P(X_n | X_1, \ldots, X_{n-1}) \\
= \quad & \ldots \\
= \quad & \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1})
\end{aligned}
$$

Simpson's paradox (Pearl, p.495)

The following table describes the effectiveness of a certain drug on a population:

	Male		Female		Overall	
	Recovered	Died	Recovered	Died	Recovered	Died
Drug used	15	40	90	50	105	90
No drug	20	40	20	10	40	50

Good news: the ratio of recovery for the whole population increases from 40/50 to 105/90

But the ratio of recovery decreases for both males and females!

Using Bayes rule for inference

Often we want to form a hypothesis about the world based on observable variables. Bayes rule is fundamental when viewed in terms of stating the belief given to a hypothesis H given evidence e:

$$P(H|e) = \frac{P(e|H)P(H)}{P(e)}$$

- $P(H|e)$ is sometimes called **posterior probability**
- $P(H)$ is called **prior probability**
- $P(e|H)$ is called **likelihood**
- $P(e)$ is just a normalizing constant, that can be computed from the requirement that $P(H|e) + P(\neg H|e) = 1$:

$$P(e) = P(e|H)P(H) + P(e|\neg H)P(\neg H)$$

Sometimes we write $P(H|e) = \alpha P(e|H)P(H)$

Example: Medical Diagnosis

A doctor knows that meningitis causes a stiff neck 80% of the time. She knows that if a person is selected randomly from the population, there is a 1/10000 chance of the person having meningitis. 1 in 100 people suffer from a stiff neck.

You go to the doctor complaining about the **symptom** of having a stiff neck. What is the probability that meningitis is the **cause** of this symptom?

Let M be meningitis, S be stiff neck:

$$P(M|S) = \frac{P(S|M)P(M)}{P(S)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

Combining predictive and diagnostic support

It is convenient to re-write Bayes rule in terms of **odds** and **likelihood ratios**:

$$\frac{P(H|e)}{P(\neg H|e)} = \frac{P(e|H)}{P(e|\neg H)} \frac{P(H)}{P(\neg H)}$$

Define the **prior odds** (predictive support) as :

$$O(H) = \frac{P(H)}{1 - P(H)} = \frac{P(H)}{1 - P(H)}$$

Define the **likelihood ratio** (diagnostic support) as:

$$L(e|H) = \frac{P(e|H)}{P(e|\neg H)}$$

Then the **posterior odds** are:

$$O(H|e) = L(e|H)O(H)$$

Computing conditional probabilities

Typically, we are interested in the posterior joint distribution of some **query variables** Y given specific values e for some **evidence variables** E

Let the **hidden variables** be $Z = X - Y - E$

If we have a joint probability distribution, we can compute the answer by “summing out” the hidden variables:

$$P(Y|e) = \alpha P(Y, e) = \alpha \sum_{h} P(Y, e, z)$$

Big problem: the joint distribution is too big to handle!

Example

Suppose we consider medical diagnosis, and there are 100 different symptoms and test results that the doctor could consider. A patient comes in complaining of fever, stiff neck and nausea. The doctor wants to compute the probability of meningitis.

- The probability table has $>= 2^{100}$ entries!
- For computing the probability of a disease, we have to sum out over 97 hidden variables!

Independence of random variables

Two random variables X and Y are **independent** (denoted $I(X,Y)$) if knowledge about X does not change the uncertainty about Y and vice versa.

$$P(X|Y) = P(X) \text{ (and vice versa)}$$

or equivalently, $P(X,Y) = P(X)P(Y)$ If n Boolean variables are independent, the whole joint distribution can be computed as:

$$P(x_1, \ldots x_n) = \prod_i P(x_i)$$

Only n numbers are needed to specify the joint, instead of 2^n

But absolute independence is a very strong requirement, seldom met

Conditional independence

Two variables X and Y are **conditionally independent** given Z if:

$$P(x|y,z) = P(x|z), \forall x, y, z$$

This means that knowing the value of Y does not change the prediction about X is the value of Z is known.

We denote this by $I(X, Y|Z)$.

Note that Pearl uses the notation $I(X, Z, Y)$

Example

Consider the dentist problem with three random variables:
$Toothache$, $Cavity$, $Catch$ (steel probe catches in my tooth)

The full joint distribution has $2^3 - 1$ = 7 independent entries

If I have a cavity, the probability that the probe catches in it does **not** depend on whether I have a toothache:

$$P(Catch|Toothache, Cavity) = P(Catch|Cavity) \tag{1}$$

I.e., $Catch$ is **conditionally independent** of $Toothache$ given $Cavity$

The same independence holds if I do not have a cavity:

$$P(Catch|Toothache, \neg Cavity) = P(Catch|\neg Cavity) \tag{2}$$

Example (continued)

Full joint distribution can now be written as:

$$
\begin{aligned}
& P(Toothache, Catch, Cavity) = \\
= \quad & P(Toothache, Catch|Cavity)P(Cavity) \\
= \quad & P(Toothache|Cavity)P(Catch|Cavity)P(Cavity)
\end{aligned}
$$

I.e., 2 + 2 + 1 = 5 independent numbers (equations 1 and 2 remove two numbers)

Much more important savings happen if the system has lots of variables!

Naive Bayesian model

A common assumption in early diagnosis is that the symptoms are independent of each other given the disease

- Let $x_1, \ldots x_n$ be the symptoms exhibited by a patient (e.g. fever, headache etc)
- Let H be the patient's health status
- Then by using the naive Bayes assumption, we get:

$$P(H, x_1, \ldots x_n) = P(H)P(x_1|H) \cdots P(x_n|H)$$

- The odds of health state given the symptoms is also easy to compute:

$$O(H|x_1, \ldots x_n) = O(H) \prod_{i=1}^{n} L(x_i|H)$$

Recursive Bayesian updating

The naive Bayes assumption allows also for a very nice, incremental updating of beliefs as more evidence is gathered

Suppose that after knowing symptoms $x_1, \ldots x_n$ the odds of H are:

$$O(H|x_1 \ldots x_n) = O(H) \prod_{i=1}^{n} L(x_i|H)$$

What happens if a new symptoms x_{n+1} appears?

$$O(H|x_1 \ldots x_n, x_{n+1}) = O(H) \prod_{i=1}^{n+1} L(x_i|H) = O(H|x_1 \ldots x_n) L(x_{n+1}|H)$$

An even nicer formula can be obtained by taking logs:

$$\log O(H|x_1 \ldots x_n, x_{n+1}) = \log O(H|x_1 \ldots x_n) + \log L(x_{n+1}|H)$$

Application: Learning to classify text

Target concept $Interesting?$: $Document \rightarrow \{+, -\}$

1. Represent each document by vector of words: one attribute per word position in document
2. Learning: Use training examples to estimate $P(+)$, $P(-)$, $P(doc|+)$, $P(doc|-)$

Naive Bayes conditional independence assumption

$$P(doc|v_j) = \prod_{i=1}^{length(doc)} P(a_i = w_k|v_j)$$

where $P(a_i = w_k|v_j)$ is probability that word in position i is w_k, given v_j

One more assumption: $P(a_i = w_k|v_j) = P(a_m = w_k|v_j), \forall i, m$

Naive Bayes Learning for Text

Input: $Examples$ (the set of documents), V (the appropriate classifications)

1. Collect all words and other tokens that occur in $Examples$ into a $Vocabulary$
2. For each target value v_j in V do
   - $docs_j$ contains the documents with target value v_j
   - $P(v_j) \leftarrow \frac{|docs_j|}{|Examples|}$
   - n is the total number of words in $docs_j$ (counting duplicate words multiple times)
   - For each word w_k in $Vocabulary$
     - n_k is the number of times word w_k occurs in $docs_j$
     - $P(w_k|v_j) \leftarrow \frac{n_k+1}{n+|Vocabulary|}$

Using the Naive Bayes Classifier

Input: a new document Doc

1. $positions \leftarrow$ all word positions in Doc that contain tokens found in $Vocabulary$
2. Return v_{NB}, where

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_{i \in positions} P(a_i | v_j)$$

Twenty NewsGroups

Given 1000 training documents from each group, learn to classify new documents according to which newsgroup they came from

comp.graphics	misc.forsale
comp.os.ms-windows.misc	rec.autos
comp.sys.ibm.pc.hardware	rec.motorcycles
comp.sys.mac.hardware	rec.sport.baseball
comp.windows.x	rec.sport.hockey
alt.atheism	sci.space
soc.religion.christian	sci.crypt
talk.religion.misc	sci.electronics
talk.politics.mideast	sci.med
talk.politics.misc	talk.politics.guns

Naive Bayes: 89% classification accuracy

For text classification, Naive Bayes obtains results comparable to any other learning algorithm

Three prisoners dilemma

Three prisoners, A, B and C have been tried for murder. One of them has been found guilty and will be executed tomorrow, the others will be released. The identity of the condemned prisoners is revealed to the guard, but not the prisoners themselves. Prisoner A calls the guard and asks: “Please give this letter to one of my friends who will be released.” The guard agrees to do it. An hour later, A calls the guard and asks whom he gave the letter to. The guard answers: “I gave it to B”.

Now A is thinking: “Before I talked to the guard, my chance of begin executed was 1/3, now it dropped to 1/2! What did I do wrong?