Understanding Probability

A series of sessions to understand the basics of probability theory through Monte Carlo simulations.

Date: 4/4/19

Session 1 (Basics of Probability)

In this session, I will motivate the need for probability in Machine learning related applications. Specifically, we will consider the following classification problem:

Motivation through classification problem

Let us assume, we have a data which contains the height and weight of some cats and dogs, i.e., for each animal, the height, weight and the corresponding label (cat/dog) is given to us. Our task is to predict the label (cat/dog) when we given the height and weight of a new animal. Due to the presence of inherent outliers, it might be difficult to assign a deterministic labelling for a given (weight, height). Hence we use a probabilistic setting and predict the probability of a given (weight, height) being a cat or dog.

Topics covered

- sample space
  - Set of all possible outcomes of a random experiment.
- Probability
  - At a high level, can be viewed as a mapping from elements in sample space to real numbers such that they sum to 1.
- Different types of sample spaces
- Finite sample space - Coin toss experiment.. {H,T} Die experiment.. {1,2,3,4,5,6}
  - Infinite but discrete
    - If you look at the experiment of tossing a coin and waiting till you see the first head, the sample space is {1,2,3,...}. In other words there are infinite possibilities, but they are discrete or countable
  - Continous sample spaces
    - If we want to sample a real number between the interval [0,1], where each number is equally likely.

- Notion of pdf (probability density function) for contionous sample spaces
  - Motivate the need to assign density function instead of probabilities for contionous sample spaces through the example of a rod with uniform density and its relation to mass of the rod.
- Explain Monte Carlo experiments
  - When we repeat the experiment a lot of times, probability can be viewed as the fraction of times, we get the desired outcome by the total number of times the experiment is conducted.

Useful Links

1. This website has some interesting visualisations for you to play with and experiment! https://seeing-theory.brown.edu/basic-probability/index.html
2. This lectures will be useful for better theoretical understanding. There are some simple exercises too to test your understanding. Try if you like to.. https://www.khanacademy.org/math/probability/probability-geometry/probability-basics/v/basic-probability
3. If you are new to Python, the following short tutorial will be useful to get started. https://www.kaggle.com/learn/python
4. This is some quick reference for functions that are useful to generate random numbers in Python. https://www.geeksforgeeks.org/random-numbers-in-python/

Date: 9/4/19

Session 2 (Conditional probability and Bayes Theorem)

The plan for the first session is to understand the following basic concepts of probability through one example problem:

Problem

Assume that there are two coins which have two faces namely head and tail. Assume that Coin 1 is fair coin, i.e., it has an equal chance of landing up with either head or tail. Assume Coin 2 is a biased coin. Specifically, let the probability of landing up on head is 0.9. With these two coins, consider the following experiment: - First, we randomly pick one of two coins to play with - We toss the picked coin 5 times. - We observed that the picked coin results in H,H,T,T,H.

Question: Coditioned on the above observations, what is the proabilility that Coin 2 was the coin that was initially picked up?

Topics Covered

- Events, Union of Events, Intersection of events
  - At a high level, understanding union as sum, intersection as multiplication and conditioning as division.
- Conditional probability
  - Understanding conditioning through monte carlo as follows. We repeat the original experiment and throw away all samples that are not meeting our condition, and caluclate the probabilities using the remaining samples only. For example, let us assume we have thrown a die and observed that an even number has occured. Now, conditioned on this observation, we have to find the probability of the die being the number 2.
- Product rule
  - Derive product rule formula using conditioning formula. For example P(A|B) will be P(A and B) by P(B)
- Total Probability Theorem
  - When we are not able to directly solve the probability of an event, we divide into small events for which we know the probabilities and use conditional probability and product rule to compute the required probability. Explain this with the example of getting one head when we pick a fair or biased coin at random. Explain the concept of going back to fundemntal sample space.
- Bayes theorem
  - Derive Bayes theorem by applying product rule twice and equating them.

Specifically, I would like to take up the following problem, and explain the above concepts while approaching to solve it:

Explaining the concept of expanding or going back to fundemental sample

We shall take a smaller version of the above problem and explain how to appropriately expand the sample space and caluclate the following probability: We are asked to choose either the biased coin or fair coin at random, and asked to toss the selected coin once. What is probability of seeing head.

We plan to first solve this problem using Monte Carlo simulations without trying to use any formulas. I would like to suggest using Python for programming. We shall certify the correctness of our answer using probability theory.

Date: 11/4/19

Session 3 (Coditing practice)

Do the following experiments using Monte carlo: - Waiting time for first head - When a point is uniformly chosen in a unit square, what is the probability that it will land up inside a unit circle (inscribed in the square). - Let us assume three coin tosses are done. We have observed that it had resulted in two heads. Conditioned on this observation, what is the probability that the second toss was a head?

Date: 16/4/19

Session 4 (Random Variables)

In this session, we try to understand the following concepts:

- Random varialbe
- Discrete random varialbe
- Continous random varialbe
- Probability distribution (PMF, Pdf)
- Independent random variable
- Motivating guassian through coffee spilling

We shall discuss about some widely used distributions like Bernouli, Binomial, Gaussian. Some distributions like Beta distribution, Dirchilet distribution which are used as prior distributions will also be introduced.

Date: 23/4/19

Session 5 (Joint distributions, Conditional distributions)

The plan is to explain the concept of joint random random variables, independence, conditional distributions and marginal distributions using the following example.

Problem

1. We are asked to generate (x,y) co-ordiantes uniformly in a given circle of radius say 1 around origin. What would be answers for the following questions?

- What is the joint distribution of X and Y?
- Are X and Y independent?
- What is the marginal distribution of X? Will it be a uniform distribution?

- What is the conditional distribution of X given Y? Is it a uniform distribution?

2. Whether the answers for the above question remain the same if we ask to generate (x,y) in a unit square around origin instead of unit cirlce?
3. How to do monte carlo simulation for the first problem of uniform points in circle?

Date: 25/4/19

Session 6 (Maximal Likelihood estimation, Maximum a posteriori estimation)

Discuss about MLE and MAP using some example. https://towardsdatascience.com/probability-concepts-explained-bayesian-inference-for-parameter-estimation-90e8930e5348

Problem

Lets say, we have some observations of a coin toss HHTH. Our aim to estimate the bias of the coin, i.e., what is the probability of head (p) of the coin.

- MLE estimation
  - We solve this problem first using MLE.
- MAP estimation
  - Next, we motivate for MAP as follows. we assume that we have some prior information about what p being used. Lets say we know that a coin which is likely to result in more tails than heads is being used. We can model it is a prior distribution on p. For example f(p)=2-2p is one such example. Then we will find the posterior probabilities and find the estimate.
- Expectation over posterior distribution
  - Instead of finding a single value of p as estimate, we look at all possible values of p and take an expectation over the posterior probabilities.

Useful Link

https://zhiyzuo.github.io/MLE-vs-MAP/#coin-toss

Date: 29/4/19

Session 7 (Coding Session on Naive Bayes classification for detecting spam messages)

Discuss about Naive Bayes and give the spam/ham dataset to classify. No in-built functions should be used!

Dataset

https://www.kaggle.com/ishansoni/sms-spam-collection-dataset

Useful Link

https://towardsdatascience.com/naive-bayes-intuition-and-implementation-ac328f9c9718

Date: 02/5/19

Session 8 (Markov Chains and Hidden Markov Model)

The following topics are covered: - Markov Property - Take the general expression for the joint distribution of a set of random variable and simplify using Markov property - Introduce the concepts of state space, transition probability matrix, steady state distribution - Motivate the concept of hidden space using Language modelling based on parts-of-speech as hidden states and actual words as observed states

Useful Link

https://www.probabilitycourse.com/chapter11/11_2_1_introduction.php

Date: 07/5/19

Session 9 (Coding session on Markov chains)

Implement a simple two state Markov chain and verify the steady state distribution using Monte Carlo simulations. The following example is used: https://medium.com/@kangeugine/hidden-markov-model-7681c22f5b9

Date: 09/5/19

Session 10 (The three different problems that we encounter in HMMs)

Identified what are some key problems to solve in the context of HMMs.

- Problem 1, Given a known model what is the likelihood of observed sequence O happening?
  - This task involves summing over all possible hidden state sequences that can potentially generate the given observed sequence.
  - A naive summation results in an exponential complexity
  - Resort to a dynamic programming algorithm which can accomplish the task in linear time.
  - The dynamic programming algorithm is called the forward algorithm.
- Problem 2, Given a known model and sequence O, what is the optimal hidden state sequence?
  - For example, given a sentence, the task of finding the parts of speech for each word in the sentence boils down to this problem when HMMs are used.
  - Viterbi algorithm can be used for this task.
- Problem 3, Given observed sequence O and number of hidden states, what is the optimal model which maximizes the probability of O?
  - Expectation-Maximization algorithm can be used for this task.

Useful links

- This blog discusses these three problems with a toy example: https://medium.com/@kangeugine/hidden-markov-model-7681c22f5b9
- These lecture slides discuss the mathematics behind these algorithms in detail: http://www.cs.cmu.edu/~aarti/Class/10701_Spring14/slides/HMM.pdf

Date: 14/5/19

Session 11 (Coding session - Supervised learning on HMMs)

We take the following labelled dataset in which a set of sentences and their corresponding parts-of-speech tags are available. The task to find the best model parameters of the HMM for this data.

Dataset https://www.kaggle.com/abhinavwalia95/how-to-loading-and-fitting-dataset-to-scikit/data

Date: 16/5/19

Session 12 (Derivation of forward algorithm for HMM)

Derived the dynamic programming algorithm for finding the likelihood of an observed sequence.

Session 13 (Conditional Random fileds and application to NLP)

We will introduce conditional random fields and apply it for for problem in Natural Language Processing called Sequence tagging or Named Entity Recognition. https://medium.com/analytics-vidhya/pos-tagging-using-conditional-random-fields-92077e5eaa31