CS 725: Foundations of Machine Learning

Overview of probability theory for ML

August 16, 2020

A review of probability theory

- Sample space(S): A sample space is defined as a set of all possible outcomes of an experiment. (For now *discrete*.)

  Example of an experiment: a coin pair toss; $S = \{HH, HT, TH, TT\}$.

- Event (E) : An event is any subset of the sample space. i.e., $E \subseteq S$.

  Often we describe events using "conditions." E.g., the event that the two coin tosses are the same, $E_{\text{same}} = \{HH, TT\}$.
  Events can be combined using logical operations on these conditions. E.g., "the two coin tosses are the same **or** the second one is T":
  $E = \{HH, TT\} \cup \{HT, TT\} = \{HH, TT, HT\}$.
- **Probability distribution,** p: A function $p : S \to [0, 1]$ s.t.
  $\forall x \in S, 0 \leq p(x) \leq 1$, $\sum_{x \in S} p(x) = 1$.

A review of probability theory

- Probability of an Event: The total weight assigned to all the outcomes in the event by a given probability distribution.

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

- Note:
  - $\Pr(S) = 1$ and $\Pr(\emptyset) = 0$
  - $\Pr(\overline{E}) = 1 - \Pr(E)$, where $\overline{E} = S \setminus E$
  - $\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2)$
  - If $E_1, E_2, \ldots, E_n$ are pairwise disjoint events, then

$$Pr(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} Pr(E_i)$$

A review of probability theory

- Conditional Probability: For events E_1, E_2 (s.t. $\Pr(E_2) > 0$), define

$$\Pr(E_1|E_2) = \frac{\Pr(E_1 \cap E_2)}{\Pr(E_2)}$$

- Bayes' Rule, named after Thomas Bayes (1701-1761):

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)},$$

provided $\Pr(A), \Pr(B) > 0$.

Using Bayes' Rule

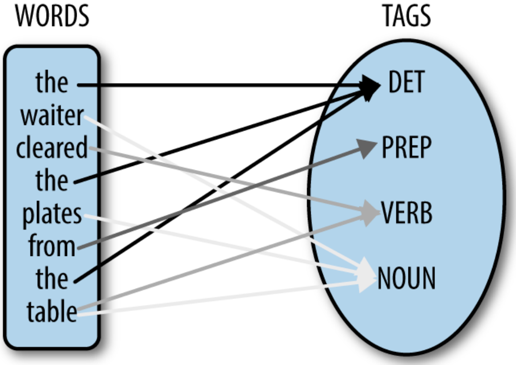
A lab test has a probability 0.95 of detecting a disease when applied to a person suffering from said disease, and a probability 0.10 of giving a false positive when applied to a non-sufferer. If 0.5% of the population are sufferers, what is the probability of a test on a random individual being positive?

Example - Part of Speech

POS tagging is a popular Natural Language Processing (**NLP**) problem

Input: A set of n-words

Output: Part-of-speech (POS) tag for each word



Assuming each word is independently drawn from a fixed vocabulary, find the probability that a sentence of length m contains a 'noun' given that it contains a 'verb'. Say the probability that a word is of POS type 'k' is p_k.

Solution:

- Let A_k be the event that the sentence contains POS type 'k'

$\Pr(A_k) = 1 - (1 - p_k)^m$

where $(1 - p_k)^m$ is the probability that all m words are not of POS type 'k'.

$\Pr(A_{noun} \mid A_{verb}) = \frac{\Pr(A_{noun} \cap A_{verb})}{\Pr(A_{verb})}$

$$\begin{aligned}
\Pr(A_j \cap A_k) &= 1 - \Pr(\overline{A_j \cap A_k}) \\
&= 1 - \Pr(\overline{A_j} \cup \overline{A_k})
\end{aligned} \tag{1}$$

We know that $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$. Now, we need $\Pr(\overline{A_j} \cap \overline{A_k})$. This refers to the event that neither tag k nor tag j are present. Thus, $\Pr(\overline{A_j} \cap \overline{A_k}) = (1 - (p_j + p_k))^m$. Using this in Equation 1, we get

$$\Pr(A_j \cap A_k) = 1 - \Pr(\overline{A_j}) - \Pr(\overline{A_k}) + (1 - (p_j + p_k))^m$$

$$= 1 - (1 - p_j)^m - (1 - p_k)^m + (1 - (p_j + p_k))^m$$

Final answer:

$Pr(A_{noun} \mid A_{verb}) = \frac{1-(1-p_{noun})^m-(1-p_{verb})^m+(1-p_{noun}-p_{verb})^m}{1-(1-p_{verb})^m}$

Random Variable

- Random variable (X) : A variable that takes values from S according to a probability distribution.
  E.g., $S = \{H, T\}$ and $\Pr(X = H) = 2/3$, $\Pr(X = T) = 1/3$.
- Jointly distributed random variables: X, Y take values from sets S_1, S_2 respectively, according to a distribution over $S_1 \times S_2$. (Generalizes to more than 2 random variables.)
  E.g., $S_1 = S_2 = \{H, T\}$ and
  $\Pr((X, Y) = (H, H)) = 1/3$, $\Pr((X, Y) = (H, T)) = 1/3$,
  $\Pr((X, Y) = (T, T)) = 1/6$, $\Pr((X, Y) = (T, H)) = 1/6$.
- Marginal Distribution: For $x \in S_1$, $\Pr(X = x) = \sum_{y \in S_2} \Pr((X, Y) = (x, y))$.
  Similarly, marginal distribution of Y can be computed.

Independence of Random Variables

- For notational convenience: $\Pr((X,Y)=(x,y))$ abbreviated as $p(x,y)$, $\Pr(X=x|Y=y)$ as $p(x|y)$ and so on.
- X,Y are said to be **independent** ($X \perp\!\!\!\perp Y$) iff for all x,y, the events $X=x$ and $Y=y$ are independent. i.e., $p(x,y)=p(x)p(y)$.
- X,Y are said to be **conditionally independent** given Z iff for all x,y,z (with $p(z)>0$), $p(x,y|z)=p(x|z)p(y|z)$.

Continuous Random Variables

Unlike discrete random variables which can take *finitely many* different values, **continuous random variables** can take on *infinitely many* values.

If the sample space for a continuous random variable is the set of all reals (i.e. $x \in \mathbb{R}$), then

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

where $p(x)$ is referred to as the **probability density function (PDF)**.

Examples of commonly used continuous distributions:

Uniform distribution: $$p(x) = \begin{cases} \frac{1}{b-a} \text{ if } a \leq x \leq b \\ 0 \text{ otherwise} \end{cases}$$

Gaussian distribution (mean μ, variance σ^2): $$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$