CS 340 Lec. 10: Probability

AD

January 2011

Sample Space and Events

- **Definition**. The *sample space* S of an experiment (whose outcome is uncertain) is the set of all possible outcomes of the experiment.
- *Example* (ranking movies): Assume Mr. X has been asked to rank 3 movies: "Karate Kid", "The Bounty Hunter" and "Citizen Kane". The outcome of the experiment is a ranking and

$$S = \{\text{all 3! permut. of "Karate Kid", "Bounty Hunter" \& "Citizen Kane"}\}.$$

- **Definition**. Any *subset* E of the sample space S is known as an *event*; i.e. an event is a set consisting of possible outcomes of the experiment.
- *Example* (ranking movies): The event $E = \{\text{all rankings in } S \text{ starting with "Citizen Kane"}\}$ is the event that Mr. X puts "Citizen Kane" at the top of his ranking.

Union and Intersection of Events

- Given events E and F, $E \cup F$ is the set of all outcomes *either* in E or F and in *both* E and F. $E \cup F$ occurs if *either* E or F occurs. $E \cup F$ is the **union** of events E and F
- *Example* (ranking movies): If we have

$$\begin{aligned} E &= \{\text{all outcomes in } S \text{ starting with ``Citizen Kane''}\}, \\ F &= \{\text{all outcomes in } S \text{ finishing with ``Karate Kid''}\} \end{aligned}$$

  then $E \cup F$ is the event that Mr. X put "Citizen Kane" at the top OR "Karate Kid" at the bottom.
- Given events E and F, $E \cap F$ is the set of all outcomes which are *both* in E and F. $E \cap F$ is also denoted EF or E, F
- *Example* (ranking movies): $E \cap F$ is the event that at the top your ranking you put "Citizen Kane" at the top and "Karate Kid" at the bottom.

Axioms of Probability

- Consider an experiment with sample space S. For each event E, we assume that a number $P(E)$, the probability of the event E, is defined and satisfies the following 3 axioms.
- **Axiom 1**
$$0 \leq P(E) \leq 1$$
- **Axiom 2**
$$P(S) = 1$$
- **Axiom 3**. For any sequence of mutually exclusive events $\{E_i\}_{i \geq 1}$, i.e. $E_i \cap E_j = \varnothing$ when $i \neq j$, then
$$P\left(\cup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

Interpretation of Probability

- Consider an event E of the sample space S. Assume you replicate the experiment n times, then it is tempting to define "practically"

$$P(E) = \lim_{n \to \infty} \frac{n(E)}{n}$$

  where $n(E)$ is the number of times the event E occurred in the n experiments.
- This is known as the frequentist approach: you should repeat an infinite number of times an experiment and the probabilities corresponds to the limiting frequencies.
- *Problem*. This kind of approach makes sense if you toss a coin but you cannot ask Mr. X one million times to rank these three movies.
- In many scenarios, probabilities are measures of the individual's degree of belief: this is *subjective*.
- This does not have any impact on the mathematical "machinery" as long as you define the axioms 1,2 and 3 are satisfied.

Conditional Probabilities and Independence

- **Conditional Probability**. Consider an experiment with sample space S. Let E and F be two events, then the conditional probability of E given F is denoted by $P(E|F)$ and satisfies if $P(F) > 0$

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{P(E, F)}{P(F)}$$

- **Intuition**: If F has occured, then, in order for E to occur, it is necessary that the occurence be a point both in E and F, hence it must be in $E \cap F$. Once F has occured, F is the new sample space.
- **Independence:** Two events E and F are said to be independent if

$$P(E, F) = P(E) P(F)$$

which implies

$$P(E|F) = P(E)$$

- *Example* (ranking movies): $E = \{$"Karate Kid" top movie for Mr. X$\}$ and $F = \{$Mr. X is a fan of martial arts$\}$ then you definitely want a probability model such that $P(E|F) \neq P(E)$.

Bayes Formula

- **Bayes Formula.** We have directly by symmetry

$$P(E|F) = \frac{P(F|E)\,P(E)}{P(F)}$$

  where

$$P(F) = P(F|E)\,P(E) + P(F|E^c)\,P(E^c).$$

- In many practical machine learning problems, you "build" $P(E, F)$ either from

$$P(E, F) = P(F|E)\,P(E)$$

  or

$$P(E, F) = P(E|F)\,P(F).$$

Conditional Independence

- We say that the events E and F are conditionally independent given G if

$$P(E, F|G) = P(E|G)\,P(F|G)\,.$$

- *Example*: Kevin separately phones two students, Alice and Bob. To each, he tells the same number; i.e. event $G = \{\text{Kevin said '7' to Alice and bob}\}$. Due to the noise in the phone, Alice and Bob each imperfectly (and independently) draw a conclusion about what number Kevin said. Let us define the events $E = \{\text{Alice heard number 7}\}$ and $F = \{\text{Bob heard number 7}\}$ respectively then E and F are conditionally independent given G as

$$P(E, F|G) = P(E|G, F)\,P(F|G) = P(E|G)\,P(F|G)$$

but we definitely expect

$$P(E|F) > P(E)$$

so the events E and F are not (marginally) independent.

- In many scenarios, we are interested in a function of the outcome as opposed to the actual outcome; e.g. we are interested in the sum of two dice and not in the separate values of each die or simply as it is easier to encode. Real-valued functions defined on the sample space are *random variables*; e.g. your score at the SAT test etc.
- A discrete r.v. X takes value in an at most countable set $\mathcal{X}$ and is defined by its *p.m.f.*

$$p_X(x) = P(X = x)$$

where

$$p_X(x) \geq 0 \text{ and } \sum_{x \in \mathcal{X}} p_X(x) = 1.$$

- *Expected value/mean and Variance*

$$\begin{aligned}
\mu &= \mathbb{E}(X) = \sum_{x \in \mathcal{X}} x \, p_X(x), \\
\mathbb{V}ar(X) &= \mathbb{E}\left((X-\mu)^2\right) = \mathbb{E}\left(X^2\right) - \mu^2.
\end{aligned}$$

Conditional Distributions: Discrete Case

- Assume X, Y are discrete-valued r.v. and take values in $\mathcal{X} \times \mathcal{Y}$ with a joint p.m.f. $p(x, y)$ then the conditional p.m.f. of X given $Y = y$ is

$$P(X = x|\, Y = y) = p_{X|Y}(x|\, y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} = \frac{p_{Y|X}(y|\, x)\, p_X(x)}{p_Y(y)}$$

  where

$$p_Y(y) = \sum_{x \in \mathcal{X}} p_{X,Y}(x, y)$$

- *Example:* $X \in \{0, 1, 2, ..., 9\}$ is a digit, Y is a 16×16 image where each pixel can take 256 values.
- *Example:* $X \in \{0, 1\}$ corresponding to spam/non spam and $Y \in \{0, 1\}^n$ is a vector of n binary variables indicating whether some prespecified words, e.g. "viagra", "money", "huge" appear in an email.

Independence and Conditional Independence of Random Variables

- Consider r.v. $X_1, X_2, ..., X_n$ with a joint p.m.f. $p_{X_1,...,X_n}(x_1, ..., x_n)$ then these variables are called independent if and only if

$$p_{X_1,...,X_n}(x_1, ..., x_n) = \prod_{i=1}^{n} p_{X_i}(x_i).$$

- Consider r.v. $X_1, X_2, ..., X_n$ with a joint p.m.f. $p_{X_1,...,X_n}(x_1, ..., x_n)$ then these variables are called independent upon Y if and only if

$$p_{X_1,...,X_n|Y}(x_1, ..., x_n|\, y) = \prod_{i=1}^{n} p_{X_i|Y}(x_i|\, y).$$

- *Example:* $Y \in \{0, 1\}$ indicates spam/non spam and $X_i \in \{0, 1\}$ indicates whether a prespecified word appears in the email.

Markov Chains

- Consider r.v. $X_1, X_2, ..., X_n$ with a joint p.m.f. $p_{X_1,...,X_n}(x_1, ..., x_n)$ then we always have

$$p_{X_1,...,X_n}(x_1, ..., x_n) = p_{X_1}(x_1) \prod_{i=2}^{n} p_{X_i | X_1,...,X_{i-1}}(x_i | x_1, ..., x_{i-1}).$$

- A sequence of r.v. $\{X_k\}_{k \geq 1}$ is said to have the Markov property if and only if

$$p_{X_i | X_1,...,X_{i-1}}(x_i | x_1, ..., x_{i-1}) = p_{X_i | X_{i-1}}(x_i | x_{i-1});$$

  i.e. the conditional distribution of X_i only depends on $(X_1, X_2, ..., X_{i-1})$ through X_{i-1}.

- Markov models are ubiquitous models for time series in Machine learning, EE, Finance etc.