Machine Learning Draws Heavily On. . .

- Probability and Statistics
- Optimization
- Algorithms and Data Structures

Probability: Foundations

A probability space $(\Omega, \mathcal{F}, P)$ consists of

- a set Ω of "possible outcomes"
- a set[1] $\mathcal{F}$ of events, which are subsets of Ω
- a probability measure $P : \mathcal{F} \to [0, 1]$ which assigns probabilities to events in $\mathcal{F}$

Example: Rolling a Die

Consider rolling a fair six-sided die. In this case,

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$
$$\mathcal{F} = \{\emptyset, \{1\}, \{2\}, \ldots, \{1, 2\}, \{1, 3\}, \ldots\}$$
$$P(\emptyset) = 0, P(\{1\}) = \frac{1}{6}, P(\{3, 6\}) = \frac{1}{3}, \ldots$$

[1] Actually, F is a σ-field. See Durrett's *Probability: Theory and Examples* for thorough coverage of the measure-theoretic basis for probability theory.

Probability: Random Variables

- A random variable is an assignment of (often numeric) values to outcomes in Ω.
- For a set A in the range of a random variable X, the induced probability that X falls in A is written as $P(X \in A)$.

Example Continued: Rolling a Die

Suppose that we bet \$5 that our die roll will yield a2. Let $X : \{1, 2, 3, 4, 5, 6\} \rightarrow \{-5, 5\}$ be a random variable denoting our winnings: $X = 5$ if the die shows 2, and $X = -5$ if not. Furthermore,

$$P(X \in \{5\}) = \frac{1}{6} \text{ and } P(X \in \{-5\}) = \frac{5}{6}.$$

Probability: Common Discrete Distributions

Common discrete distributions for a random variable X:

- Bernoulli(p): $p \in [0, 1]$; $X \in \{0, 1\}$

$$P(X = 1) = p, \; P(X = 0) = 1 - p$$

- Binomial(p, n): $p \in [0, 1]$, $n \in \mathbb{N}$; $X \in \{0, \ldots, n\}$

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

- The multinomial distribution generalizes the Bernoulli and the Binomial beyond binary outcomes for individual experiments.
- Poisson(λ): $\lambda \in (0, \infty)$; $X \in \mathbb{N}$

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Probability: More on Random Variables

- Notation: $X \sim P$ means "X has the distribution given by P"
- The cumulative distribution function (cdf) of a random variable $X \in \mathbb{R}^m$ is defined for $x \in \mathbb{R}^m$ as $F(x) = P(X \leq x)$.
- We say that X has a density function p if we can write $P(X \leq x) = \int_{-\infty}^{x} p(y)dy$.
- In practice, the continuous random variables with which we will work will have densities.
- For convenience, in the remainder of this lecture we will assume that all random variables take values in some countable numeric set, $\mathbb{R}$, or a real vector space.

Probability: Common Continuous Distributions

Common continuous distributions for a random variable X:

- Uniform(a, b): $a, b \in \mathbb{R}$, $a < b$; $X \in [a, b]$

$$p(x) = \frac{1}{b - a}$$

- Normal(μ, σ^2): $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_{++}$; $X \in \mathbb{R}$

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

- Normal distribution can be easily generalized to the multivariate case, in which $X \in \mathbb{R}^m$. In this context, μ becomes a real vector and σ is replaced by a covariance matrix.
- Beta, Gamma, and Dirichlet distributions also frequently arise.

Exponential Family

- encompasses distributions of the form

$$P(X = x) = h(x)\exp(\eta(\theta)T(x) - A(\theta))$$

- includes many commonly encountered distributions
- well-studied and has various nice analytical properties while being fairly general

Graphical Models

Graphical models provide a flexible framework for building complex models involving many random variables while allowing us to leverage conditional independence relationships among them to control computational tractability.