Fundamentals

Stat 241A/CS 281A: Statistical Learning Theory

Hongwei Li

Based on tutorial slides by Yangqing Jia, Po-Ling Loh, Lester Mackey and Ariel Kleiner

August 29, 2012

Outline

Probability: Foundations

Definition

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

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

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

Example: Rolling a Dice

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

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

$$\mathcal{F} = \{\emptyset, \{1\}, \{2\}, \ldots, \{1, 2\}, \{1, 3\}, \ldots\}$$

$$P(\emptyset) = 0, P(\{1\}) = \frac{1}{6}, P(\{3, 6\}) = \frac{1}{3}, \ldots$$

Probability: Random Variables

Definition

A **random variable** X is an assignment of (often numeric) values to outcomes ω in the sample space Ω

- X is a function of the sample space (e.g., $X : \Omega \to \mathbb{R}$)
- We write $P(X \in A)$ to mean the induced probability that the value of X falls in a set A
  - Formally, $P(X \in A) \triangleq P(\{\omega \in \Omega : X(\omega) \in A\})$
- $X \sim P$ means "X has the distribution given by P"

Example Continued: Rolling a Die

Suppose that we bet \$5 that our die roll will yield a 2.
Let X be a random variable denoting our winnings:

- $X : \Omega = \{1, 2, 3, 4, 5, 6\} \to \{-5, 5\}$
- $X = 5$ if the die shows 2, and $X = -5$ if not
- $P(X \in \{5\}) = \frac{1}{6}$ and $P(X \in \{-5\}) = \frac{5}{6}$.

Probability: Common Discrete Distributions

Common discrete distributions for a random variable X:

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

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

e.g., X = 1 if biased coin comes up heads, 0 otherwise

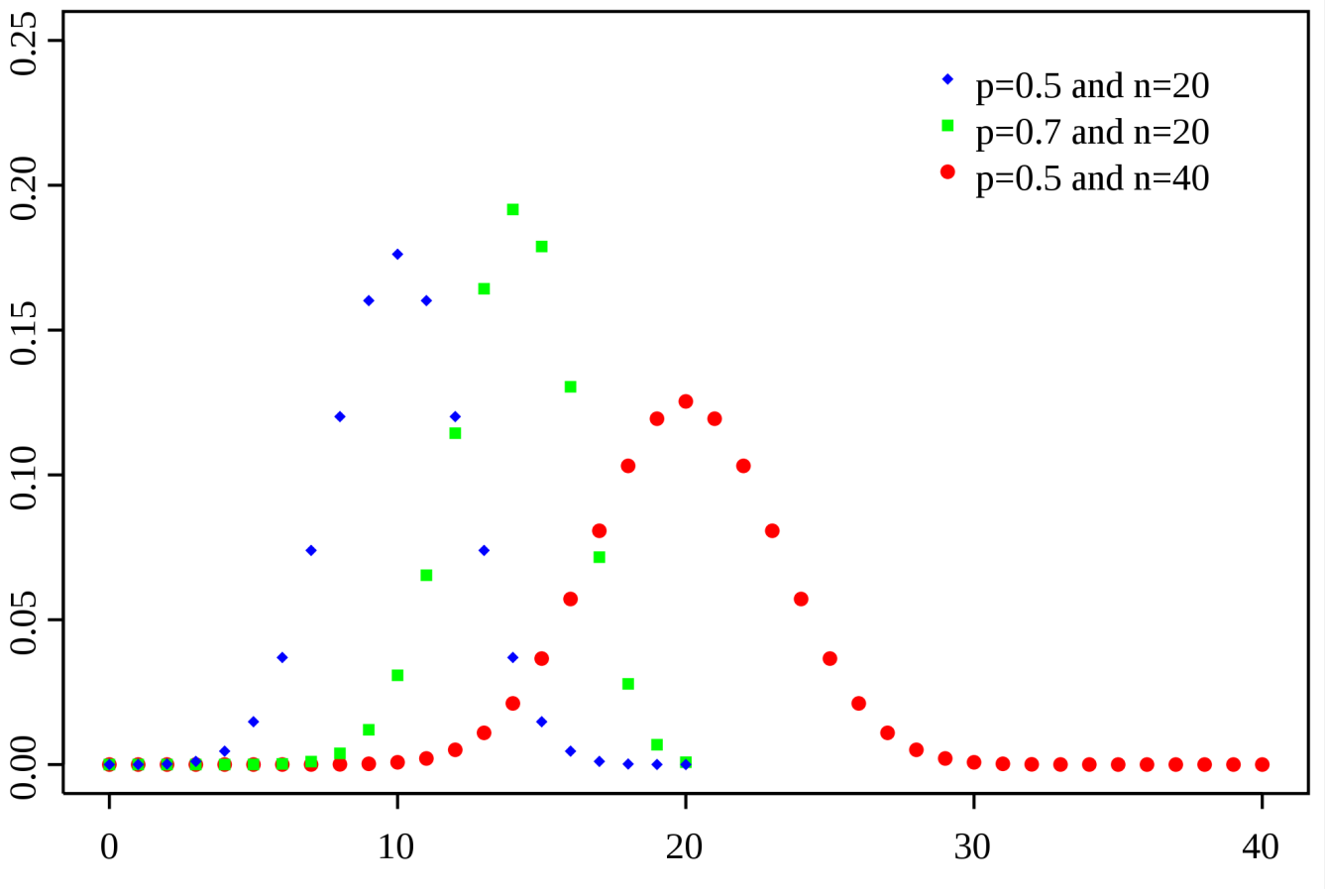
Probability: Common Discrete Distributions

Common discrete distributions for a random variable X:

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

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

e.g., X = number of heads in n tosses of a biased coin



Probability: Common Discrete Distributions

Common discrete distributions for a random variable X:

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

$$P(X = x) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

  - Generalizes Bernoulli and Binomial to non-binary outcomes
  - $\mathbf{p}$ is a vector of probabilities summing to 1
  - X is a vector of counts summing to n

  e.g., X = number of times each digit rolled in n rolls of a die

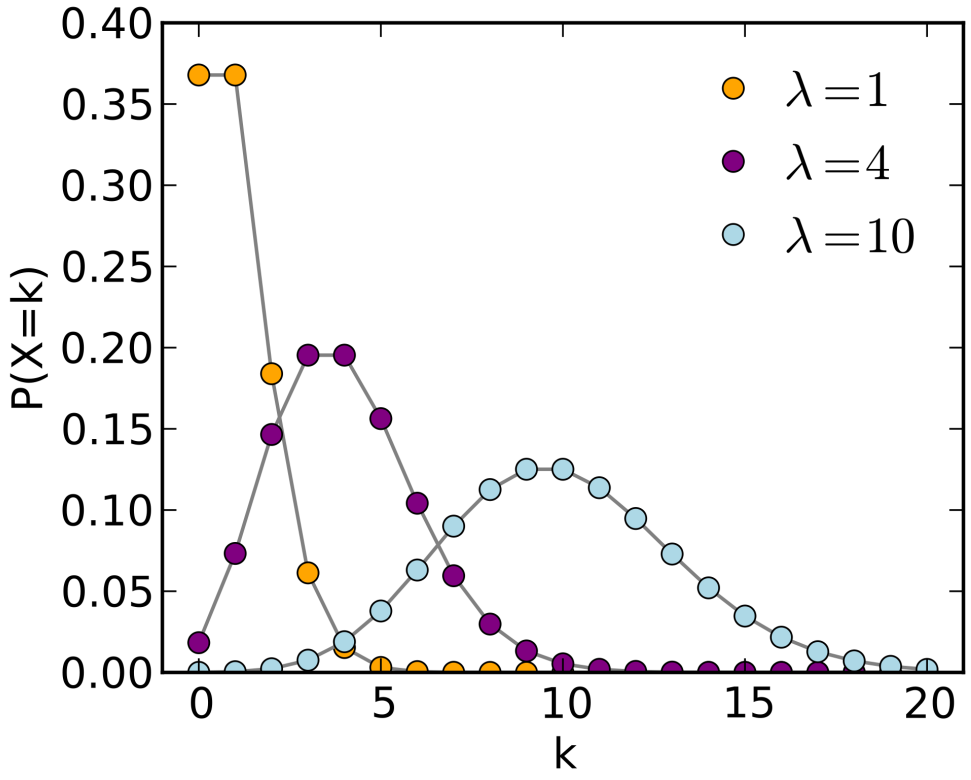
Probability: Common Discrete Distributions

Common discrete distributions for a random variable X:

- Poisson(λ): $\lambda \in (0, \infty)$; $X \in \mathbb{N}$

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

e.g., X = number of deaths by horse-kicking each year



Probability: From Discrete to Continuous

Definition

The **probability mass function** (pmf) of a discrete random variable X is defined as $p(x) = P(X = x)$.

Definition

The **cumulative distribution function** (cdf) of a random variable $X \in \mathbb{R}^m$ is defined for $x \in \mathbb{R}^m$ as $F(x) = P(X \leq x)$.

Definition

We say that X has a **probability density function** (pdf) p if we can write $F(x) = \int_{-\infty}^{x} p(y)dy$.

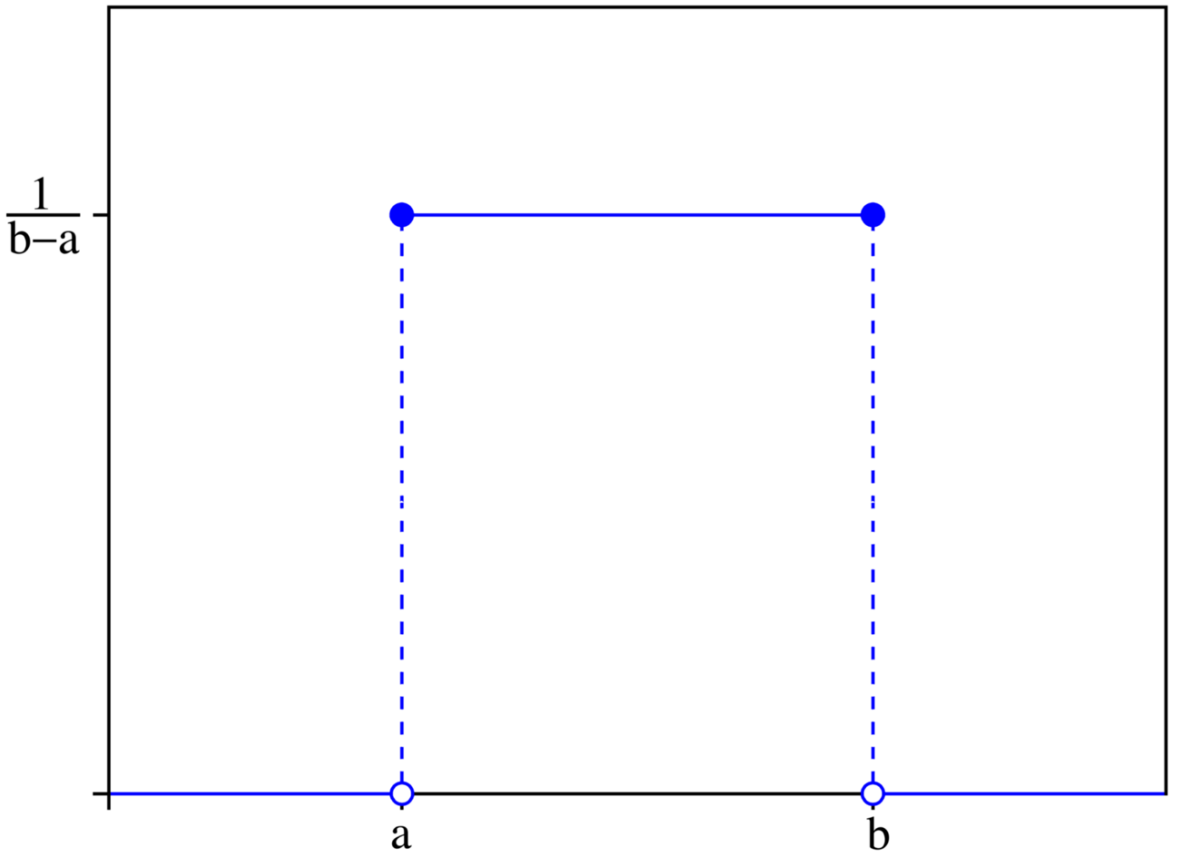
- In practice, the continuous random variables with which we will work will have densities.
- For convenience, in the remainder of this lecture we will assume that all random variables take values in some countable numeric set, $\mathbb{R}$, or a real vector space.

Probability: Common Continuous Distributions

Common continuous distributions for a random variable X:

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

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

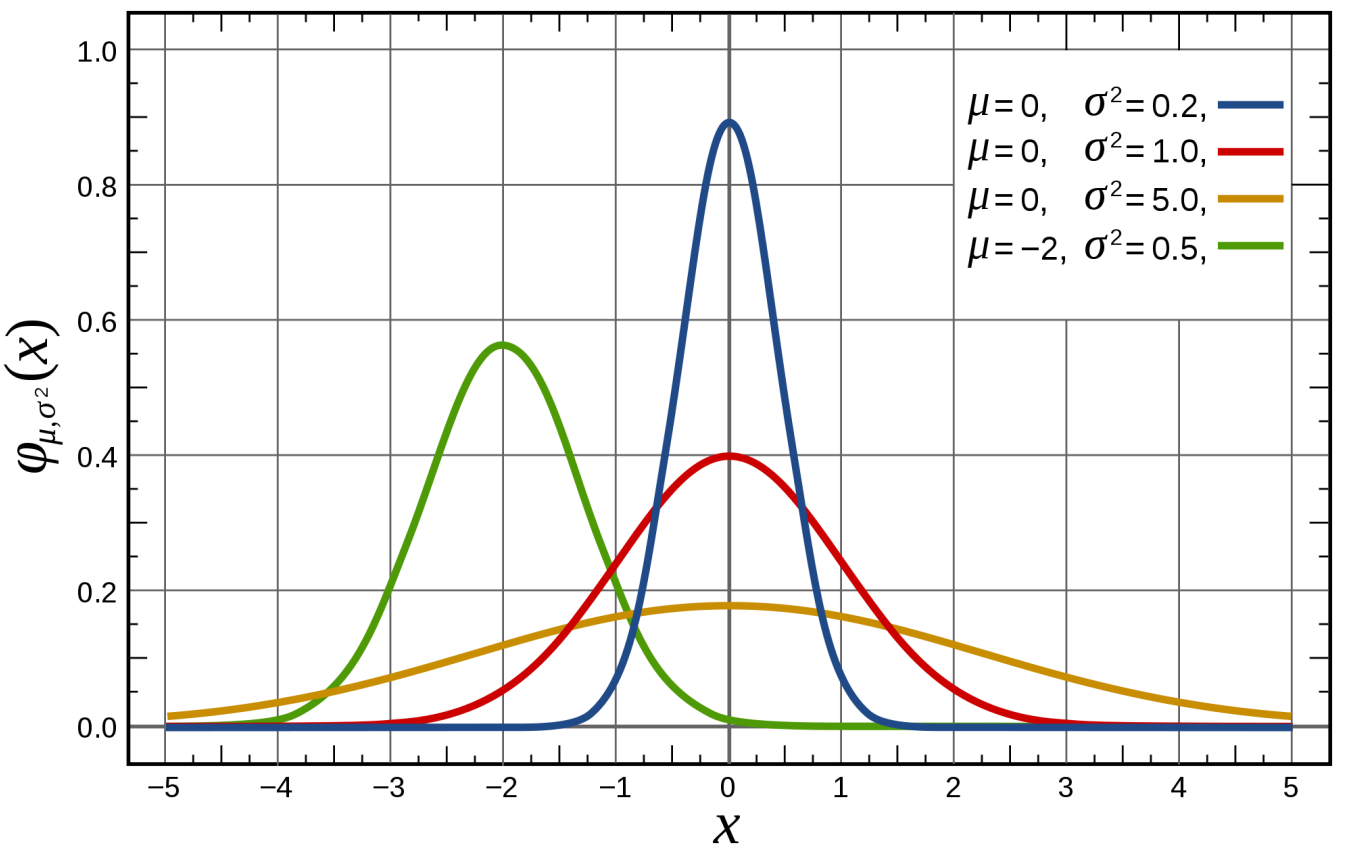


Probability: Common Continuous Distributions

Common continuous distributions for a random variable X:

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

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

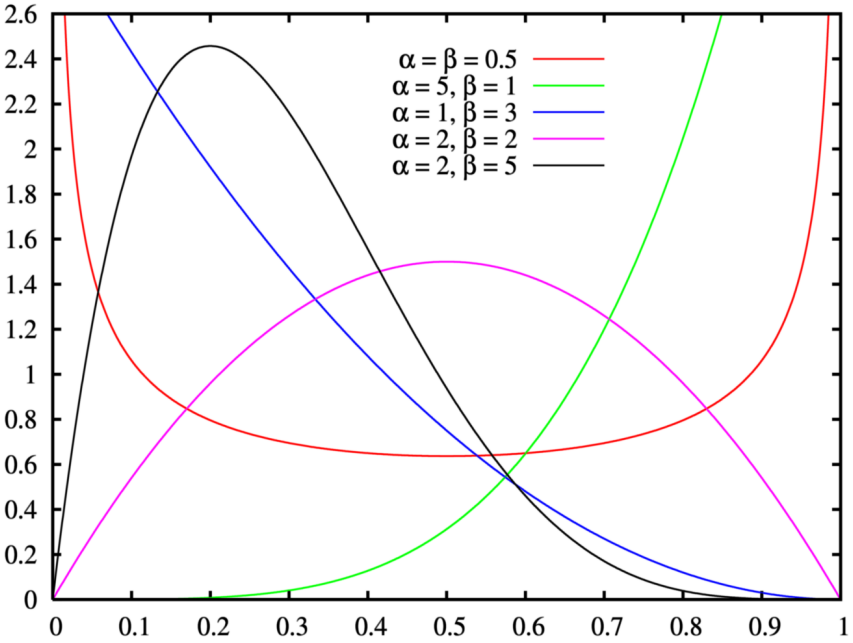


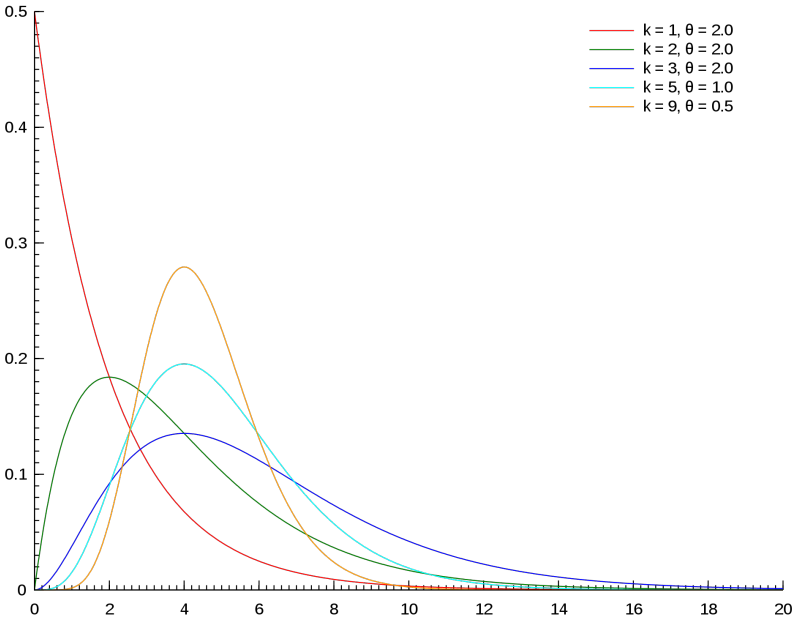
- Normal distribution can be easily generalized to the

Probability: Common Continuous Distributions

Common continuous distributions for a random variable X:

- Beta, Gamma, and Dirichlet distributions also frequently arise.





Exponential Family

- Encompasses distributions of the form

$$p(x) = h(x)\exp(\eta(\theta)^T T(x) - A(\theta))$$

- Well-studied, nice analytical properties
- Includes many commonly encountered distributions
  - Binomial(p,n): for fixed n and varying parameter p

$$\begin{aligned} P(X = x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \binom{n}{x} \exp\left(x \log\left(\frac{p}{1-p} \right) + n \log(1-p) \right) \end{aligned}$$

  - Bernoulli, Multinomial, Normal, Poisson, ...

Probability: Expectation

Intuition: the expectation of random variable is its "average" value under its distribution

Definition

Formally, the **expectation** of a random variable X, denoted $E[X]$, is its integral with respect to its probability measure P.

- If X takes values in some countable numeric set $\mathcal{X}$, then

$$E[X] = \sum_{x \in \mathcal{X}} xP(X = x)$$

- If $X \in \mathbb{R}^m$ has a density p, then

$$E[X] = \int_{\mathbb{R}^m} xp(x)dx$$

Probability: More on Expectation

Properties of Expectation

- Expection is linear: $E[aX + b] = aE[X] + b$. Also, if Y is also a random variable, then $E[X + Y] = E[X] + E[Y]$.
- Expectation is monotone: if $X \geq Y$, then $E[X] \geq E[Y]$
- Probabilities are expectations:
  - Let $\mathbf{1}_A$ equal 1 when the event A occurs and 0 otherwise
  - $E[\mathbf{1}_A] = P(\mathbf{1}_A = 1)1 + P(\mathbf{1}_A = 0)0 = P(\mathbf{1}_A = 1) = P(A)$
- Expectations also obey various inequalities, including Jensen's, Cauchy-Schwarz, etc.

Variance

The variance of a random variable X is defined as

$$\text{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

and obeys the following for $a, b \in \mathbb{R}$:

$$\text{Var}(aX + b) = a^2 \, Var(X).$$

Probability: Independence

Intuition: two random variables are independent if knowing the value of one yields no knowledge about the value of the other

Definition

Formally, two random variables X and Y are **independent**, written $X \perp Y$, iff

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

for all (measurable) subsets A and B in the ranges of X and Y.

- If X, Y have densities $p_X(x), p_Y(y)$, then they are independent if

  $$p_{X,Y}(x, y) = p_X(x)p_Y(y)$$

  for all x, y.

Probability: Conditioning

Intuition: conditioning allows us to capture the probabilistic relationships between different random variables

Definition

For events A and $B \in \mathcal{F}$, $P(A|B)$ is the probability that A will occur given that we know that event B has occurred.

- If $P(B) > 0$, then $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$

Example: Random variables X and Y

- $P(X \in C | Y \in D) = \dfrac{P(X \in C, Y \in D)}{P(Y \in D)}$
- In terms of densities, $p(y|x) = \dfrac{p(x,y)}{p(x)}$, for $p(x) > 0$ where $p(x) = \int p(x,y)dy$.
- If X and Y are independent, $P(X \in C | Y \in D) = P(X \in C)$.

Probability: More on Conditional Probability

For any events A and B (e.g., we might have $A = \{Y \leq 5\}$),

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

Bayes' Theorem

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

Equivalently, if $P(B) > 0$, $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$

- Bayes' Theorem provides a means of inverting the "order" of conditioning

Probability: Conditional Independence

Intuition: conditioning can induce independence

Definition

Formally, two random variables X and Y are **conditionally independent** given a third random variable Z, written $X \perp Y|Z$, iff

$$P(X \in A, Y \in B|Z = z) = P(X \in A|Z = z)P(Y \in B|Z = z)$$

for all (measurable) subsets A and B in the ranges of X and Y and all values z in the range of Z.

- In terms of densities, $X \perp Y|Z$ if

$$p_{X,Y|Z}(x, y|z) = p_{X|Z}(x|z)p_{Y|Z}(y|z)$$

for all x, y, z.

Statistics: Frequentist Basics

Given: Data $x_1, x_2, \ldots, x_n$

- Realizations of random variables, $X_1, \ldots, X_n$, generally assumed independent and identically distributed (i.i.d.)

Goal: Estimate a *parameter* θ

- Some (unknown) value associated with the distribution generating the data
- Our estimate will be a *statistic*, i.e., a function $\hat{\theta}(x_1, \ldots, x_n)$ of the data

Examples

- Given the results of n independent flips of a coin, determine the probability p with which it lands on heads.
- Or, simply determine whether or not the coin is fair.
- Find a function that distinguishes digital images of fives from those of other handwritten digits.

Statistics: Parameter Estimation

Important Question: How do we estimate θ?

- Generally, θ indexes a class of probability distributions: $\{p_\theta(x) : \theta \in \Theta\}$
- How do we choose $\hat{\theta}(x_1, \ldots, x_n)$ so that $p_{\hat{\theta}}(x)$ best reflects our data?
- One answer: **maximize the likelihood** (or, equivalently, log likelihood) of the data
  - $\ell(\theta; x_1, \ldots, x_n) = p_\theta(x_1, \ldots, x_n) = \prod_{i=1}^n p_\theta(x_i)$
  - $\ln \ell(\theta; x_1, \ldots, x_n) = \sum_{i=1}^n \ln p_\theta(x_i)$

Maximum Likelihood Estimation

$$\hat{\theta}(x_1, \ldots, x_n) = \operatorname*{argmax}_{\theta \in \Theta} \prod_{i=1}^n p_\theta(x_i) = \operatorname*{argmax}_{\theta \in \Theta} \sum_{i=1}^n \ln p_\theta(x_i)$$

Statistics: Maximum Likelihood Estimation

Example: Normal Mean

- Suppose that our data $x_1, \ldots, x_n$ is real-valued and known to be drawn i.i.d. from a normal distribution with variance 1 but unknown mean.
- **Goal**: estimate the mean θ of the distribution.
- Recall that a univariate $N(\theta, 1)$ distribution has density $p_\theta(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(x-\theta)^2)$.
- Given data $x_1, \ldots, x_n$, we can obtain the maximum likelihood estimate by maximizing the log likelihood w.r.t. θ:

$$\frac{d}{d\theta} \sum_{i=1}^{n} \ln p_\theta(x_i) \propto \sum_{i=1}^{n} \frac{d}{d\theta} \left[-\frac{1}{2}(x_i - \theta)^2 \right] = \sum_{i=1}^{n} (x_i - \theta) = 0$$

$$\Rightarrow \hat{\theta}(x_1, \ldots, x_n) = \operatorname*{argmax}_{\theta \in \Theta} \sum_{i=1}^{n} \ln p_\theta(x_i) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Statistics: Bayesian Basics

- The Bayesian approach treats parameters as random variables having distributions.
- That is, we maintain probability distributions over possible parameter values:
  1. We have some beliefs about our parameter values θ before we see any data. These beliefs are encoded in the **prior distribution** $p(\theta)$.
  2. Treating the parameters θ as random variables, we can write the likelihood of the data $X = x$ as a conditional probability: $p(x|\theta)$.
  3. We would like to update our beliefs about θ based on the data by obtaining $p(\theta|x)$, the **posterior distribution**. **Solution**: by Bayes' theorem,

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

where

$$p(x) = \int p(x|\theta)p(\theta)d\theta$$

Statistics: More on the Bayesian Approach

- Within the Bayesian framework, estimation and prediction simply reduce to probabilistic inference. This inference can, however, be analytically and computationally challenging.
- It is possible to obtain point estimates from the posterior in various ways, such as by taking the posterior mean

$$E_{\theta|X}[\theta] = \int \theta p(\theta|x) d\theta$$

or the mode of the posterior:

$$\operatorname*{argmax}_{\theta} p(\theta|x)$$

- Alternatively, we can directly compute the predictive distribution of a new data point X_{new}, having already seen data $X = x$:

$$p(x_{\text{new}}|x) = \int p(x_{\text{new}}|\theta) p(\theta|x) d\theta$$

Statistics: Bayesian Approach for the Normal Mean

Suppose that $X|\theta \sim N(\theta, 1)$ and we place a prior $N(0, 1)$ over θ (i.e., $\theta \sim N(0, 1)$):

$$p_{X|\theta}(x|\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right) \qquad p_\theta(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\theta^2}{2}\right)$$

Then, if we observe X = 1,

$$\begin{aligned} p_{\theta|X}(\theta|1) &= \frac{p_{X|\theta}(1|\theta)p_\theta(\theta)}{p_X(1)} \\ &\propto p_{X|\theta}(1|\theta)p_\theta(\theta) \\ &= \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(1-\theta)^2}{2}\right)\right]\left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\theta^2}{2}\right)\right] \\ &\propto \frac{1}{.5\sqrt{2\pi}} \exp\left(-\frac{(\theta-.5)^2}{2(.5)}\right) = N(0.5, 0.5) \end{aligned}$$

Statistics: Bayesian Prior Distributions

Important Question: How do we select our prior distribution?

Different possible approaches:

- Based on actual prior knowledge about the system or data generation mechanism
- Target analytical and computational tractability; e.g., use conjugate priors (those which yield posterior distributions in the same family)
- Allow the data to have "maximal impact" on the posterior

Statistics: Parametric vs. Non-Parametric Models

- All of the models considered so far are **parametric** models: they are determined by a fixed, finite number of parameters.
- This can limit the flexibility of the model.
- Instead, can permit a potentially infinite number of parameters which is allowed to grow as we see more data. Such models are called **non-parametric**.
- Although non-parametric models yield greater modeling flexibility, they are generally statistically and computationally less efficient.

Statistics: Generative vs. Discriminative Models

- Suppose that, based on data $(x_1, y_1), \ldots, (x_n, y_n)$, we would like to obtain a model whereby we can predict the value of Y based on an always-observed random variable X.
- **Generative Approach**: model the full joint distribution $P(X, Y)$, which fully characterizes the relationship between the random variables.
- **Discriminative Approach**: only model the conditional distribution $P(Y|X)$
- Both approaches have strengths and weaknesses and are useful in different contexts.

Linear Algebra: Basics

Matrix Transpose

- For an $m \times n$ matrix A with $(A)_{ij} = a_{ij}$, its transpose is an $n \times m$ matrix A^T with $(A^T)_{ij} = a_{ji}$.
- $(AB)^T = B^T A^T$

Matrix Inverse

- The inverse of a square matrix $A \in \mathbb{R}^{n \times n}$ is the matrix A^{-1} such that $A^{-1}A = I$.
- This notion generalizes to non-square matrices via left- and right-inverses.
- Not all matrices have inverses.
- If A and B are invertible, then $(AB)^{-1} = B^{-1}A^{-1}$.
- Computation of inverses generally requires $O(n^3)$ time.

Linear Algebra: Basics

Trace

- For a square matrix $A \in \mathbb{R}^{n \times n}$, its trace is defined as $\text{tr}(A) = \sum_{i=1}^{n} (A)_{ii}$.
- $\text{tr}(AB) = \text{tr}(BA)$

Eigenvectors and Eigenvalues

- Given a matrix $A \in \mathbb{R}^{n \times n}$, $u \in \mathbb{R}^n \backslash \{0\}$ is called an eigenvector of A with $\lambda \in \mathbb{R}$ the corresponding eigenvalue if

$$Au = \lambda u$$

- An $n \times n$ matrix can have no more than n distinct eigenvector/eigenvalue pairs.

Linear Algebra: Basics

More definitions

- A matrix A is called *symmetric* if it is square and $(A)_{ij} = (A)_{ji}, \forall i, j$.
- A symmetric matrix A is *positive semi-definite (PSD)* if all of its eigenvalues are greater than or equal to 0.
- Changing the above inequality to $>$, $\leq$, or $<$ yields the definitions of positive definite, negative semi-definite, and negative definite matrices, respectively.
- A positive definite matrix is guaranteed to have an inverse.

Linear Algebra: Matrix Decompositions

Eigenvalue Decomposition

Any symmetric matrix $A \in \mathbb{R}^{n \times n}$ can be decomposed as follows:

$$A = U \Lambda U^T$$

where Λ is a diagonal matrix with the eigenvalues of A on its diagonal, U has the corresponding eigenvectors of A as its columns, and $UU^T = I$.

Singular Value Decomposition

Any matrix $A \in \mathbb{R}^{m \times n}$ can be decomposed as follows:

$$A = U \Sigma V^T$$

where $UU^T = VV^T = I$ and Σ is diagonal.

Other Decompositions: LU (into lower and upper triangular matrices); QR; Cholesky (only for PSD matrices)

Optimization: Basics

- We often seek to find optima (minima or maxima) of some real-valued vector function $f : \mathbb{R}^n \to \mathbb{R}$. For example, we might have $f(x) = x^T x$.
- Furthermore, we often constrain the value of x in some way: for example, we might require that $x \geq 0$.
- In standard notation, we write

$$\begin{aligned} \min_{x \in \mathcal{X}} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0, i = 1, \ldots, N \\ & h_i(x) = 0, i = 1, \ldots, M \end{aligned}$$

- Every such problem has a (frequently useful) corresponding Lagrange dual problem which lower-bounds the original, primal problem and, under certain conditions, has the same solution.
- It is only possible to solve these optimization problems analytically in special cases, though we can often find solutions numerically.

Optimization: A Simple Example

- Consider the following unconstrained optimization problem:

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 = \min_{x \in \mathbb{R}^n} (Ax - b)^T (Ax - b)$$

- In fact, this is the optimization problem that we must solve to perform least-squares regression.
- To solve it, we can simply set the gradient of the objective function equal to 0.
- The gradient of a function $f(x) : \mathbb{R}^n \to \mathbb{R}$ is the vector of partial derivatives with respect to the components of x:

$$\nabla_x f(x) = \left(\frac{\partial f}{\partial x_1}, \ldots \frac{\partial f}{\partial x_n} \right)$$

Optimization: A Simple Example

Thus, we have

$$\begin{aligned}
\nabla_x \|Ax - b\|_2^2 &= \nabla_x \left[(Ax - b)^T (Ax - b) \right] \\
&= \nabla_x \left[x^T A^T A x - 2x^T A^T b + b^T b \right] \\
&= 2A^T Ax - 2A^T b \\
&= 0
\end{aligned}$$

and so the solution is

$$x = (A^T A)^{-1} A^T b$$

(if $(A^T A)^{-1}$ exists).

Optimization: Convexity

- In the previous example, we were guaranteed to obtain a global minimum because the objective function was *convex*.
- A twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if its Hessian (matrix of second derivatives) is everywhere PSD (if $n = 1$, then this corresponds to the second derivative being everywhere non-negative)[1].
- An optimization problem is called convex if its objective function f and inequality constraint functions $g_1, \ldots, g_N$ are all convex, and its equality constraint functions $h_1, \ldots, h_M$ are linear.
- For a convex problem, all minima are in fact global minima. In practice, we can efficiently compute minima for problems in a number of large, useful classes of convex problems.

[1] This definition is in fact a special case of the general definition for arbitrary vector functions.