Chapter 2: Basics from Probability Theory and Statistics

2.1 Probability Theory

Events, Probabilities, Random Variables, Distributions, Moments

Generating Functions, Deviation Bounds, Limit Theorems

Basics from Information Theory

2.2 Statistical Inference: Sampling and Estimation

Moment Estimation, Confidence Intervals

Parameter Estimation, Maximum Likelihood, EM Iteration

2.3 Statistical Inference: Hypothesis Testing and Regression

Statistical Tests, p-Values, Chi-Square Test

Linear and Logistic Regression

mostly following L. Wasserman Chapters 1-5, with additions from other textbooks on stochastics

2.1 Basic Probability Theory

A **probability space** is a triple (Ω, E, P) with

- a set Ω of elementary events (sample space),
- a family E of subsets of Ω with $\Omega \in E$ which is closed under $\cap$, $\cup$, and $-$ with a countable number of operands (with finite Ω usually $E=2^{\Omega}$), and
- a **probability measure $P: E \rightarrow [0,1]$** with $P[\Omega]=1$ and $P[\cup_i A_i] = \sum_i P[A_i]$ for countably many, pairwise disjoint A_i

Properties of P:

$P[A] + P[\neg A] = 1$

$P[A \cup B] = P[A] + P[B] - P[A \cap B]$

$P[\emptyset] = 0$ (null/impossible event)

$P[\Omega\ ] = 1$ (true/certain event)

Independence and Conditional Probabilities

Two events A, B of a prob. space are **independent** if $P[A \cap B] = P[A]\, P[B]$.

A finite set of events $A=\{A_1, ..., A_n\}$ is **independent** if for every subset $S \subseteq A$ the equation $P[\bigcap_{A_i \in S} A_i] = \prod_{A_i \in S} P[A_i]$ holds.

The **conditional probability P[A | B]** of A under the condition (hypothesis) B is defined as: $$P[A \mid B] = \frac{P[A \cap B]}{P[B]}$$

Event A is **conditionally independent** of B given C if $P[A \mid BC] = P[A \mid C]$.

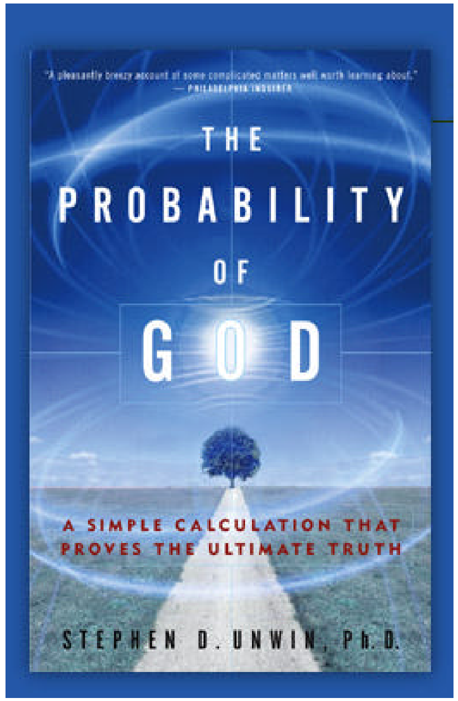
Total Probability and Bayes' Theorem

Total probability theorem:

For a partitioning of Ω into events $B_1, ..., B_n$:

$$P[A] = \sum_{i=1}^{n} P[A \mid B_i]\, P[B_i]$$

Bayes' theorem:
$$P[A \mid B] = \frac{P[B \mid A]\, P[A]}{P[B]}$$



P[A|B] is called ***posterior probability***

P[A] is called ***prior probability***

Random Variables

A **random variable (RV)** X on the prob. space (Ω, E, P) is a function $X: \Omega \rightarrow M$ with $M \subseteq R$ s.t. $\{e \mid X(e) \leq x\} \in E$ for all $x \in M$
(X is measurable).

$F_X: M \rightarrow [0,1]$ with $F_X(x) = P[X \leq x]$ is the ***(cumulative) distribution function (cdf)*** of X.
With countable set M the function $f_X: M \rightarrow [0,1]$ with $f_X(x) = P[X = x]$ is called the ***(probability) density function (pdf)*** of X;
in general $f_X(x)$ is $F'_X(x)$.

For a random variable X with distribution function F, the inverse function $F^{-1}(q) := \inf\{x \mid F(x) > q\}$ for $q \in [0,1]$ is called ***quantile function*** of X.
(0.5 quantile (50^{th} percentile) is called median)

Random variables with countable M are called ***discrete***, otherwise they are called ***continuous***.
For discrete random variables the density function is also referred to as the ***probability mass function***.

Important Discrete Distributions

- **Bernoulli** distribution with parameter p: $P[X = x] = p^x (1-p)^{1-x}$ for $x \in \{0,1\}$
- **Uniform** distribution over {1, 2, ..., m}:

$$P[X=k] = f_X(k) = \frac{1}{m} \quad for\, 1 \le k \le m$$

- **Binomial** distribution (coin toss n times repeated; X: #heads):

$$P[X=k] = f_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- **Poisson** distribution (with rate λ):

$$P[X=k] = f_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

- **Geometric** distribution (#coin tosses until first head):

$$P[X=k] = f_X(k) = (1-p)^k \, p$$

- **2-Poisson mixture** (with $a_1+a_2=1$):

$$P[X=k] = f_X(k) = a_1 e^{-\lambda_1} \frac{\lambda_1^{\,k}}{k!} + a_2 e^{-\lambda_2} \frac{\lambda_2^{\,k}}{k!}$$

Important Continuous Distributions

- **Uniform** distribution in the interval [a,b]

$$f_X(x) = \frac{1}{b-a} \quad \text{for } a \le x \le b \ (0 \text{ otherwise})$$

- **Exponential** distribution (z.B. time until next event of a Poisson process) with rate $\lambda = \lim_{\Delta t \to 0}$ (# events in Δt) / Δt :

$$f_X(x) = \lambda e^{-\lambda x} \quad \text{for } x \ge 0 \ (0 \text{ otherwise})$$

- **Hyperexponential** distribution: $f_X(x) = p\lambda_1 e^{-\lambda_1 x} + (1-p)\lambda_2 e^{-\lambda_2 x}$

- **Pareto** distribution: $f_X(x) \to \frac{a}{b}\left(\frac{b}{x}\right)^{a+1} \quad \text{for } x > b, 0 \text{ otherwise}$

  Example of a „heavy-tailed“ distribution with $f_X(x) \to \frac{c}{x^{\alpha+1}}$

- **logistic** distribution: $F_X(x) = \frac{1}{1+e^{-x}}$

Normal Distribution (Gaussian Distribution)

- ***Normal distribution N(μ,σ^2)*** (Gauss distribution; approximates sums of independent, identically distributed random variables):

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Distribution function of N(0,1):

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

<u>Theorem:</u>

Let X be normal distributed with expectation μ and variance σ^2.

Then $$Y := \frac{X-\mu}{\sigma}$$

is normal distributed with expectation 0 and variance 1.

Multidimensional (Multivariate) Distributions

Let $X_1, ..., X_m$ be random variables over the same prob. space with domains $dom(X_1), ..., dom(X_m)$.

The ***joint distribution*** of $X_1, ..., X_m$ has a density function

$$f_{X_1,...,X_m}(x_1,...,x_m)$$

$$with \sum_{x_1 \in dom(X_1)} ... \sum_{x_m \in dom(X_m)} f_{X_1,...,X_m}(x_1,...,x_m) = 1$$

$$or \int_{dom(X_1)} ... \int_{dom(X_m)} f_{X1,...,Xm}(x_1,...,x_m)\, dx_m ... dx_1 = 1$$

The ***marginal distribution*** of X_i in the joint distribution of $X_1, ..., X_m$ has the density function

$$\sum_{x_1} ... \sum_{x_{i-1}} \sum_{x_{i+1}} ... \sum_{x_m} f_{X_1,...,X_m}(x_1,...,x_m) \ or$$

$$\int_{X_1} ... \int_{X_{i-1}} \int_{X_{i+1}} ... \int_{X_m} f_{X_1,...,X_m}(x_1,...,x_m)\, dx_m \ ... \ dx_{i+1}\ dx_{i-1} ... dx_1$$

Important Multivariate Distributions

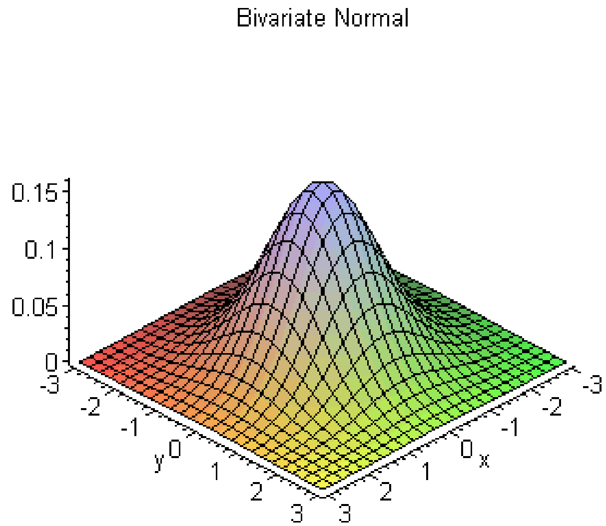
multinomial distribution (n trials with m-sided dice):

$$P[X_1 = k_1 \wedge \ldots \wedge X_m = k_m] = f_{X_1,\ldots,X_m}(k_1,\ldots,k_m) = \binom{n}{k_1 \ldots k_m} p_1^{k_1} \ldots p_m^{k_m}$$

$$\text{with } \binom{n}{k_1 \ldots k_m} := \frac{n!}{k_1! \ldots k_m!}$$

multidimensional normal distribution:

$$f_{X_1,\ldots,X_m}(\vec{x}) = \frac{1}{\sqrt{(2\pi)^m |\Sigma|}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})}$$



with covariance matrix Σ with $\Sigma_{ij} := Cov(X_i, X_j)$

Moments

For a discrete random variable X with density f_X

$$E[X] = \sum_{k \in M} k\, f_X(k)$$ is the ***expectation value (mean)*** of X

$$E[X^i] = \sum_{k \in M} k^i\, f_X(k)$$ is the ***i-th moment*** of X

$$V[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$ is the ***variance*** of X

For a continuous random variable X with density f_X

$$E[X] = \int_{-\infty}^{+\infty} x\, f_X(x)\, dx$$ is the ***expectation value*** of X

$$E[X^i] = \int_{-\infty}^{+\infty} x^i\, f_X(x)\, dx$$ is the ***i-th moment*** of X

$$V[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$ is the ***variance*** of X

Theorem: Expectation values are additive: $E[X+Y] = E[X] + E[Y]$
(distributions are not)

Properties of Expectation and Variance

$E[aX+b] = aE[X]+b$ for constants a, b

$E[X_1+X_2+...+X_n] = E[X_1] + E[X_2] + ... + E[X_n]$
(i.e. expectation values are generally additive, but distributions are not!)

$E[X_1+X_2+...+X_N] = E[N]\ E[X]$
if $X_1, X_2, ..., X_N$ are independent and identically distributed **(iid RVs)** with mean $E[X]$ and N is a stopping-time RV

$Var[aX+b] = a^2\ Var[X]$ for constants a, b

$Var[X_1+X_2+...+X_n] = Var[X_1] + Var[X_2] + ... + Var[X_n]$
if $X_1, X_2, ..., X_n$ are independent RVs

$Var[X_1+X_2+...+X_N] = E[N]\ Var[X] + E[X]^2\ Var[N]$
if $X_1, X_2, ..., X_N$ are iid RVs with mean $E[X]$ and variance $Var[X]$ and N is a stopping-time RV

Correlation of Random Variables

Covariance of random variables Xi and Xj::

$$Cov(Xi, Xj) := E[(Xi - E[Xi])(Xj - E[Xj])]$$

$$Var(Xi) = Cov(Xi, Xi) = E[X^2] - E[X]^2$$

Correlation coefficient of Xi and Xj

$$\rho(Xi, Xj) := \frac{Cov(Xi, Xj)}{\sqrt{Var(Xi)}\sqrt{Var(Xj)}}$$

Conditional expectation of X given Y=y:

$$E[X \mid Y = y] = \begin{cases} \sum x\, f_{X|Y}(x \mid y) & \text{discrete case} \\ \int x\, f_{X|Y}(x \mid y)\, dx & \text{continuous case} \end{cases}$$

Transformations of Random Variables

Consider expressions r(X,Y) over RVs such as X+Y, max(X,Y), etc.

1. For each z find $A_z = \{(x,y) \mid r(x,y) \le z\}$
2. Find cdf $F_Z(z) = P[r(x,y) \le z] = \int\int_{A_Z} f_{X,Y}(x,y)\,dx\,dy$
3. Find pdf $f_Z(z) = F'_Z(z)$

Important case: ***sum of independent RVs*** (non-negative)

$Z = X+Y$

$$F_Z(z) = P[r(x,y) \le z] = \int_y\int_x{}_{x+y\le z} f_X(x) f_Y(y)\,dx\,dy$$

$$= \int_{y=0}^{z-x}\int_{x=0}^{z} f_X(x) f_Y(y)\,dx\,dy$$

$$= \int_{x=0}^{z} f_X(x) F_Y(z-x)\,dx$$

or in discrete case:

Convolution

$$F_Z(z) = \sum_x \sum_y{}_{x+y\le z} f_X(x) f_Y(y)$$

Generating Functions and Transforms

X, Y, ...: continuous random variables with non-negative real values

A, B, ...: discrete random variables with non-negative integer values

$$M_X(s) = \int_0^\infty e^{sx} f_X(x)dx = E[e^{sX}]:$$

moment-generating function of X

$$f^*_X(s) = \int_0^\infty e^{-sx} f_X(x)dx = E[e^{-sX}]$$

Laplace-Stieltjes transform (LST) of X

$$G_A(z) = \sum_{i=0}^\infty z^i f_A(i) = E[z^A]:$$

generating function of A (z transform)

$$f_A^*(-s) = M_A(s) = G_A(e^s)$$

Examples:

exponential:

$$f_X(x) = \alpha e^{-\alpha x}$$

$$f^*_X(s) = \frac{\alpha}{\alpha + s}$$

Erlang-k:

$$f_X(x) = \frac{\alpha k(\alpha k x)^{k-1}}{(k-1)!} e^{-\alpha k x}$$

$$f^*_X(s) = \left(\frac{k\alpha}{k\alpha + s}\right)^k$$

Poisson:

$$f_A(k) = e^{-\alpha} \frac{\alpha^k}{k!}$$

$$G_A(z) = e^{\alpha(z-1)}$$

Properties of Transforms

$$M_X(s) = 1 + sE[X] + \frac{s^2 E[X^2]}{2!} + \frac{s^3 E[X^3]}{3!} + \ldots \qquad f_A(n) = \frac{1}{n!} \frac{d^n G_A(z)}{dz^n}(0)$$

$$\Rightarrow E[X^n] = \frac{d^n M_X(s)}{ds^n}(0) \qquad E[A] = \frac{dG_A(z)}{dz}(1)$$

$$f_X(x) = ag(x) + bh(x) \Rightarrow f^*(s) = ag^*(s) + bh^*(s)$$

$$f_X(x) = g'(x) \Rightarrow f^*(s) = sg^*(s) - g(0^-)$$

$$f_X(x) = \int_0^x g(t)dt \Rightarrow f^*(s) = \frac{g^*(s)}{s}$$

Convolution of independent random variables:

$$F_{X+Y}(z) = \int_0^z f_X(x) F_Y(z-x)\,dx \qquad F_{A+B}(k) = \sum_{i=o}^{k} f_A(i) F_Y(k-i)$$

$$f^*_{X+Y}(s) = f^*_X(s)\, f^*_Y(s)$$

$$M_{X+Y}(s) = M_X(s) M_Y(s) \qquad G_{A+B}(z) = G_A(z) G_B(z)$$

Inequalities and Tail Bounds

Markov inequality: $P[X \geq t] \leq E[X] / t$ for $t > 0$ and non-neg. RV X

Chebyshev inequality: $P[\,|X - E[X]| \geq t] \leq Var[X] / t^2$
for $t > 0$ and non-neg. RV X

Chernoff-Hoeffding bound: $P[X \geq t] \leq \inf\left\{ e^{-\theta t} M_X(\theta) \mid \theta \geq 0 \right\}$

Corollary: $:P\left[\left|\frac{1}{n}\sum X_i - p\right| \geq t\right] \leq 2e^{-2nt^2}$ for Bernoulli(p) iid. RVs $X_1, ..., X_n$ and any $t > 0$

Mill's inequality: $P\left[|Z| > t\right] \leq \frac{\sqrt{2}}{\pi} \frac{e^{-t^2/2}}{t}$ for N(0,1) distr. RV Z and $t > 0$

Cauchy-Schwarz inequality: $E[XY] \leq \sqrt{E[X^2]E[Y^2]}$

Jensen's inequality: $E[g(X)] \geq g(E[X])$ for convex function g
$E[g(X)] \leq g(E[X])$ for concave function g

(g is convex if for all $c \in [0,1]$ and x_1, x_2: $g(cx_1 + (1-c)x_2) \leq cg(x_1) + (1-c)g(x_2)$)

Convergence of Random Variables

Let $X_1, X_2, \ldots$ be a sequence of RVs with cdf's $F_1, F_2, \ldots,$ and let X be another RV with cdf F.

- X_n ***converges*** to X ***in probability***, $X_n \rightarrow_P X$, if for every $\varepsilon > 0$
  $P[|X_n - X| > \varepsilon] \rightarrow 0$ as $n \rightarrow \infty$
- X_n ***converges*** to X ***in distribution***, $X_n \rightarrow_D X$, if
  $\lim_{n\rightarrow\infty} F_n(x) = F(x)$ at all x for which F is continuous
- X_n ***converges*** to X ***in quadratic mean***, $X_n \rightarrow_{qm} X$, if
  $E[(X_n - X)^2] \rightarrow 0$ as $n \rightarrow \infty$
- X_n ***converges*** to X ***almost surely***, $X_n \rightarrow_{as} X$, if $P[X_n \rightarrow X] = 1$

weak law of large numbers (for $\bar{X}_n = \sum_{i=1..n} X_i / n$)
if $X_1, X_2, \ldots, X_n, \ldots$ are iid RVs with mean E[X], then $\bar{X}_n \rightarrow_P E[X]$
that is: $\lim_{n\rightarrow\infty} P[|\bar{X}_n - E[X]| > \varepsilon] = 0$

strong law of large numbers:
if $X1, X_2, \ldots, X_n, \ldots$ are iid RVs with mean E[X], then $\bar{X}_n \rightarrow_{as} E[X]$
that is: $P[\lim_{n\rightarrow\infty} |\bar{X}_n - E[X]| > \varepsilon] = 0$

Poisson Approximates Binomial

<u>Theorem:</u>

Let X be a random variable with binomial distribution with parameters n and p := α/n with large n and small constant α << 1.

Then $\lim_{n \to \infty} f_X(k) = e^{-\alpha} \frac{\alpha^k}{k!}$

Central Limit Theorem

Theorem:

Let $X_1, \ldots, X_n$ be independent, identically distributed random variables with expectation μ and variance σ^2.
The distribution function Fn of the random variable $Z_n := X_1 + \ldots + X_n$ converges to a normal distribution $N(n\mu, n\sigma^2)$ with expectation $n\mu$ and variance $n\sigma^2$:

$$lim_{n \to \infty} P[\, a \le \frac{Z_n - n\mu}{\sqrt{n}\sigma} \le b \,] = \Phi(b) - \Phi(a)$$

Corollary:

$\overline{X} := \frac{1}{n} \sum_{i=1}^{n} X_i$ converges to a normal distribution $N(\mu, \sigma^2/n)$ with expectation μ and variance σ^2/n .

Elementary Information Theory

Let f(x) be the probability (or relative frequency) of the x-th symbol in some text d. The **entropy** of the text (or the underlying prob. distribution f) is:

$$H(d) = \sum_x f(x) \log_2 \frac{1}{f(x)}$$

H(d) is a lower bound for the bits per symbol needed with optimal coding (compression).

For two prob. distributions f(x) and g(x) the **relative entropy (Kullback-Leibler divergence)** of f to g is

$$D(f \,\|\, g) := \sum_x f(x) \log \frac{f(x)}{g(x)}$$

Relative entropy is a measure for the (dis-)similarity of two probability or frequency distributions.
It corresponds to the average number of additional bits needed for coding information (events) with distribution f when using an optimal code for distribution g.

The **cross entropy** of f(x) to g(x) is:

$$H(f, g) := H(f) + D(f \,\|\, g) = -\sum_x f(x) \log g(x)$$

Compression

- Text is sequence of symbols (with specific frequencies)
- Symbols can be
  - letters or other characters from some alphabet Σ
  - strings of fixed length (e.g. trigrams)
  - or words, bits, syllables, phrases, etc.

Limits of compression:

Let p_i be the probability (or relative frequency) of the i-th symbol in text d

Then the ***entropy*** of the text:

$$H(d) = \sum_i p_i \log_2 \frac{1}{p_i}$$

is a ***lower bound*** for the average number of bits per symbol in any compression (e.g. Huffman codes)

Note:

compression schemes such as *Ziv-Lempel* (used in zip) are better because they consider context beyond single symbols; with appropriately generalized notions of entropy the lower-bound theorem does still hold