Probability Reference Sheet

By Danny Liu

1 Basic Knowledge

Let E be an event and S the sample space, then $P(E) \geq 0$, $P(S) = 1$, $P(E) = 1 - P(\overline{E})$
If $A \perp B$, then $P(A \cap B) = P(A)P(B)$ and $P(A|B) = P(A)$, If $A \cap B = \varnothing$, then $P(A \cap B) = 0$
$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$ and $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
If $S = B_1 \cup ... \cup B_n$ and $A \subseteq S$, then $P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i)$
$p(x_1, \ldots, x_n) = p(x_n|x_{n-1}, \ldots, x_1)p(x_{n-1}|x_{n-2}, \ldots, x_1) \ldots p(x_1)$

2 Bayesian Inference

$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}$, where $p(D|\theta)$ is the likelihood, $p(\theta)$ is the prior and $p(D)$ is the marginal

2.1 Conjugate Priors

A family F of priors $p(\theta)$ is conjugate to the likelihood $p(D|\theta)$ if the posterior $p(\theta|D)$ is in F
$\mathrm{Beta}(\alpha, \beta)$ is conjugate to $\mathrm{Bin}(n, p)$, $\mathrm{Geo}(p)$, $\mathrm{Bern}(\phi)$. $\mathrm{Dir}(\boldsymbol{\alpha})$ is conjugate to $\mathrm{Cat}(\boldsymbol{p})$, $\mathrm{Mult}(n, \boldsymbol{p})$
$\mathcal{N}(\mu, \sigma^2)$ is self-conjugate. $\mathrm{Gamma}(\alpha, \beta)$ is conjugate to $\mathrm{Exp}(\lambda)$, $\mathrm{Poi}(\lambda)$, $\mathcal{N}(\mu, \sigma^2)$

2.2 Maximum Likelihood Estimate

Find θ that maximizes the likelihood, $\theta^* = \underset{\theta}{\mathrm{argmax}}\, p(D|\theta)$. Assuming that D is iid, then
The conditional likelihood of $y|x$ is $L(\theta) = \prod_{i=1}^{n} p(y^{(i)}|x^{(i)}; \theta)$

2.3 Maximum a Posteriori

Given prior $p(\theta)$, find θ that maximizes the posterior, $\theta^* = \underset{\theta}{\mathrm{argmax}}\, p(\theta|D) = \underset{\theta}{\mathrm{argmax}}\, p(D|\theta)p(\theta)$
When $n \to \infty$ or $p(\theta)$ is uniform, then MAP = MLE as $p(\theta|D) \propto p(D|\theta)$

3 Expectation and Variance

$E[X] = \int_x xf(x)dx$, $\mathrm{var}(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$, $\sigma^2 = \frac{\sum(X-\mu)}{N}$
$E[g(X)] = \int_x g(x)f(x)dx$, $E[aX + b] = aE[X] + b$, $\mathrm{var}(aX + b) = a^2\mathrm{var}(X)$
$E[\sum_i X_i] = \sum_i E[X_i]$, $\mathrm{var}(\sum_i X_i) = \sum_i \mathrm{var}(X_i) + 2\sum_{i<j} \mathrm{cov}(X_i, X_j)$
$\mathrm{cov}(X, Y) = \langle X, Y\rangle = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$
Define correlation of X and Y as $\rho(X, Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}} = \frac{\langle X,Y\rangle}{\|X\|\|Y\|} = \cos\theta$
$\rho(X, Y) = \pm 1 \iff Y = aX + b$ for some $a, b \in \mathbb{F}$

3.1 Covariance Matrix

$\Sigma = E[(x-\mu)(x-\mu)^T] = E[xx^T] - \mu\mu^T$, where $\Sigma_{ij} = \text{cov}(x_i, x_j)$ and $\text{cov}(Ax+b) = A\Sigma A^T$
Given data $\{x^{(i)}\}_{i=1}^n$ with 0 mean, the empirical $\Sigma = \frac{1}{n}\sum_{i=1}^n x^{(i)}(x^{(i)})^T = \frac{1}{n}X^TX$
$\Sigma_{ij} = \Sigma_{ji}$ for $i \neq j$ are the covariances and linearly correlate the ith and jth dimensions
The diagonal entries of Σ are the variances along the original axes whereas the eigenvalues are the variances along the principal axes, hence $\Sigma \succeq 0$ as we can always rotate to the principal axes

3.2 Conditional Expectation

The conditional distribution of X given Y is $p(x|y) = \frac{p(x,y)}{p(y)} = \frac{P[X=x \cap Y=y]}{P[Y=y]}$
The conditional expectation $E[X|Y] = E_{p(x|y)}[X] = \sum_x xp(x|y)$ is a function of y
$E[E[X|Y]] = \sum_y \sum_x xp(x|y)p(y) = E[X]$, $\sum_x p(x|y) = 1$, $\sum_y p(x|y) = p(x)$
Suppose $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ is jointly Gaussian with parameters $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, then
The marginals are $p(x_1) = \mathcal{N}(x_1|\mu_1, \Sigma_{11})$ and $p(x_2) = \mathcal{N}(x_2|\mu_2, \Sigma_{22})$ with posterior conditional $p(x_1|x_2) = \mathcal{N}(x_1|\mu_{1|2}, \Sigma_{1|2})$, where $\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$ and $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

4 Inequalities and Limit Theorems

If X only takes on non-negative values, then for any $a > 0$, $P[X \geq a] \leq \frac{E[X]}{a}$ (Markov)
If $E[X] = \mu$ and $\text{var}(X) = \sigma^2$, then for any $k > 0$, $P[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$ (Chebyshev)
If f is a convex function, then $E[fX] \geq f(EX)$ (Jensen)
Let $X_1, X_2, ..., X_n$ be iid with $E[X] = \mu$ and $\text{var}(X) = \sigma^2 < \infty$, then for n sufficiently large

1. $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \approx Z \sim \mathcal{N}(0,1)$, where $\bar{X} = \frac{X_1+X_2+...+X_n}{n}$ has mean μ and variance $\frac{\sigma^2}{n}$ (CLT)
2. $P\left[\lim\limits_{n\to\infty} \frac{X_1+X_2+...+X_n}{n} = \mu\right] = 1$ (Law of Large Numbers)

5 Moment Generating Functions

$M_X(t) = E[e^{tX}] = \int_x e^{tx}f(x)dx$, where $E[X], E[X^2], E[X^3], ...$ are the moments of X
Consider Taylor expansion of $e^{tX} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + ...$
If two r.vs have the same mgf for all $t \in \mathbb{R}$, then they have the same distribution for all $t \in \mathbb{R}$.
$M_X(t) = \sum_{k=0}^\infty E[x^k]\frac{t^k}{k!}$ and $E[X^k] = \frac{d^k}{dt^k}M_X(t)\big|_{t=0}$ for any integer $k > 0$
If X, Y are r.vs and $Y = aX + b$, then $M_Y(t) = e^{bt}M_X(at)$
If $X_1, X_2, ..., X_n$ are independent r.vs, then $M_{X_1+X_2+...+X_n}(t) = M_{X_1}(t)M_{X_2}(t)...M_{X_n}(t)$

6 Random Variables

A random variable is a function $X : S \to \mathbb{R}$ where S is the sample space.
For discrete X, Y, the joint pmf is $p(x,y) = P[X = x \cap Y = y]$, where $\sum_x \sum_y p(x,y) = 1$
For continuous X, Y, the joint pdf is $\int\int_{(x,y)\in R} f(x,y)dxdy = P[(x,y) \in R]$
Note that although $\int_a^b f(x)dx = P[a \leq X \leq b]$ for continuous X, $P[X = a] = 0$
If $F(a) = P[X \leq a] = \int_{-\infty}^a f(x)dx$ is the cdf, then $f(x) = \frac{d}{dx}F(x)$
Continuous (discrete) r.vs are independent if and only if $f(x,y) = f(x)f(y)$
The marginal density of X is $f(x) = \int_y f(x,y)dy$

6.1 Functions of Random Variables

Suppose X is a random variable and $Y = g(X)$. If g is increasing and differentiable, then
$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \cdot \frac{d}{dy}g^{-1}(y) & \text{if } g(a) \leq y \leq g(b) \\ 0 & \text{otherwise} \end{cases}$

6.2 Continuity Correction

If a discrete distribution (e.g Binomial, Poisson) is approximated by a continuous distribution (e.g normal), then the following adjustments must be made
$x = n$ becomes $n - 0.5 < x < n + 0.5$
$x > n$ becomes $x > n + 0.5$ and $x < n$ becomes $x < n - 0.5$

7 Common Distributions

7.1 Bernoulli Distribution

$X \sim \text{Bernoulli}(\phi)$, $p(x) = \phi^x(1-\phi)^x$, $M_X(t) = (1-\phi) + \phi e^t$, $E[X] = \phi$, $\text{var}(X) = \phi(1-\phi)$

7.2 Categorial Distribution

$X \sim \text{Cat}(\boldsymbol{p})$, $p(\boldsymbol{x}) = \prod\limits_{i=1}^{k} p_i^{x_i}$ is a generalization of the Bernoulli/special case of multinomial

7.3 Binomial Distribution

$X \sim \text{Bin}(n, p)$ models the number of n Bernoulli trials that end up being successes
$p(x) = p^x(1-p)^{n-x}\binom{n}{x}$, $M_X(t) = (pe^t + 1 - p)^n$, $E[X] = np$, $\text{var}(X) = np(1-p)$

7.4 Multinomial Distribution

$X \sim \text{Multinomial}(n, \boldsymbol{p})$ models the outcome of n categorial trials
$p(\boldsymbol{x}) = \binom{n}{x_1,\ldots,x_k} \prod\limits_{i=1}^{k} p_i^{x_i}$, $M_X(\boldsymbol{t}) = (\sum\limits_{i=1}^{k} p_i e^{t_i})^n$, where $\sum_i p_i = 1$ and $\sum_i x_i = n$

7.5 Geometric Distribution

$X \sim \text{Geo}(p)$ models the number of Bernoulli trials needed to get one success
$p(x) = (1-p)^{x-1}p$, $M_X(t) = \frac{pe^t}{1-(1-p)e^t}$, $E[X] = \frac{1}{p}$, $\text{var}(X) = \frac{1-p}{p^2}$

7.6 Poisson Distribution

$X \sim \text{Poi}(\lambda)$ models the number of times a given event occurs independently in a fixed interval
$p(x) = e^{-\lambda}\frac{\lambda^x}{x!}$, $M_X(t) = e^{\lambda(e^t-1)}$, $E[X] = \text{var}(X) = \lambda =$ rate at which event occurs
Approximates binomial well when n is large, p is small and $\lambda = np$ is moderate

7.7 Uniform Distribution

$X \sim U(a, b)$, $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$, $M_X(t) = \frac{e^{tb}-e^{ta}}{t(b-a)}$, $E[X] = \frac{1}{2}(a+b)$, $\text{var}(X) = \frac{(b-a)^2}{12}$

7.8 Exponential Distribution

$X \sim \text{Exp}(\lambda)$ models the time between events in a Poisson process with rate λ
$f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, $M_X(t) = \frac{\lambda}{\lambda - t}$, $E[X] = \frac{1}{\lambda}$, $\text{var}(X) = \frac{1}{\lambda^2}$
Key property memoryless and continuous analogue of the geometric distribution

7.9 Gamma Distribution

$X \sim \text{Gamma}(\alpha, \beta)$ models the amount of time until α events occur in a Poisson process with rate β
$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$, $M_X(t) = (\frac{\beta}{\beta - t})^\alpha$, $E[X] = \frac{\alpha}{\beta}$, $\text{var}(X) = \frac{\alpha}{\beta^2}$

7.10 Gaussian Distribution

$X \sim \mathcal{N}(\mu, \sigma^2)$, $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(\frac{-(x-\mu)^2}{2\sigma^2})$, $M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$, $E[X] = \mu$, $\text{var}(X) = \sigma^2$
$P(a \leq X \leq b) = P(Z \leq \frac{b-\mu}{\sigma}) - P(Z \leq \frac{a-\mu}{\sigma})$, where $Z \sim \mathcal{N}(0, 1)$
Approximates binomial well (with continuity correction) when np and $n(1-p) > 5$

7.11 Multivariate Gaussian

$X \sim \mathcal{N}(\mu, \Sigma)$, where $f(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$
If the joint $f(x)$ is radially symmetric, then $\Sigma_{ij} = 0$ for $i \neq j$ and $\Sigma_{ii} = \Sigma_{jj}$ for $i = j$

7.12 Laplace Distribution

$X \sim \text{Laplace}(\mu, b)$, $f(x) = \frac{1}{2b} \exp(-\frac{|x-\mu|}{b})$, $M_X(t) = \frac{\exp(\mu t)}{1 - b^2 t^2}$, $E[X] = \mu$, $\text{var}(X) = 2b^2$
If $X, Y \sim \text{Exp}(\lambda)$, then $X - Y \sim \text{Lap}(0, \lambda^{-1})$. If $X \sim \text{Lap}(\mu, b)$, then $kX + c \sim \text{Lap}(k\mu + c, kb)$
Laplace has a fatter tail and taller head than the Gaussian due to abs. value and constant term

7.13 Beta Distribution

$X \sim \text{Beta}(\alpha, \beta)$ represents a distribution of probabilities of an uncertain binomial variable
$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}$, where $x \in [0, 1]$, $\alpha, \beta > 0$ and $E[X] = \frac{\alpha}{\alpha + \beta}$
Note $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function and $\Gamma(n) = (n-1)!$ is the gamma function
Well suited to represent a prior, as the posterior after n Bernoulli trials is $\text{Beta}(\alpha + \text{ success}, \beta + \text{ fail})$

7.14 Dirichlet Distribution

$X \sim \text{Dir}(\boldsymbol{\alpha})$ represents a distribution of probabilities of an uncertain multinomial variable
$f(\boldsymbol{x}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$, where $x_i \in [0, 1]$, $\sum_i x_i = 1$, $\alpha_i > 0$ and $E[X_i] = \frac{\alpha_i}{\sum_k \alpha_k}$

8 Calculus

Let f be a continuous real-valued function on $[a, b]$ and F be the antiderivative of f
Then $\int_a^b f(t)dt = F(b) - F(a)$ and $\frac{d}{dx} \int_a^x f(t)dt = f(x)$
$\int u dv = uv - \int v du$, $\int_s^\infty \lambda e^{-\lambda x} dx = e^{-\lambda s}$, $\int x e^{-\lambda x} dx = \frac{1}{\lambda^2}(-\lambda e^{-\lambda x} x - e^{-\lambda x}) + C$
$\int \frac{1}{1+x^2} dx = \arctan(x) + C$, $\int \frac{x^2}{1+x^2} dx = \int 1 - \frac{1}{1+x^2} dx$, $\int \frac{x^3}{1+x^2} dx = \int x - \frac{x}{1+x^2} dx$