CS838-1 Advanced NLP: Mathematical Background

Xiaojin Zhu

2007
Send comments to jerryzhu@cs.wisc.edu

These are 'just enough' as a quick reference for our course. For details, please consult any standard text book.

1 Probability

The probability of a discrete random variable A taking the value a is $P(A = a) \in [0, 1]$. Sometimes written as $P(a)$ when no danger of confusion.

Normalization $\sum_{\text{all } a} P(A = a) = 1$.

Joint probability $P(A = a, B = b) = P(a, b)$, the two events both happen at the same time.

Marginalization $P(A = a) = \sum_{\text{all } b} P(A = a, B = b)$, "summing out B".

Conditional probability $P(a|b) = \frac{P(a,b)}{P(b)}$, a happens given b.

The product rule $P(a, b) = P(a)P(b|a) = P(b)P(a|b)$.

Bayes rule $P(a|b) = \frac{P(b|a)P(a)}{P(b)}$. In general $P(a|b, C) = \frac{P(b|a,C)P(a|C)}{P(b|C)}$ where C can be one or more random variables. In the special case when θ is model parameter, D is observed data, we have

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)},$$

where $p(\theta)$ is called the prior, $p(D|\theta)$ the likelihood function (of θ, *not normalized*: $\int p(D|\theta)\, d\theta \neq 1$), $p(D)$ the evidence, and $p(\theta|D)$ the posterior.

Independence: iff A and B are independent, the product rule can be simplified as $P(a, b) = P(a)P(b)$. Equivalently, $P(a|b) = P(a)$, $P(b|a) = P(b)$.

A continuous random variable x has a probability density function (pdf) $p(x) \in [0, \infty]$.

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} p(x)\, dx$$

$$\int_{-\infty}^{\infty} p(x)\, dx = 1$$

$$p(x) = \int_{-\infty}^{\infty} p(x, y)\, dy$$

The expectation ("mean" or "average") of a function f under the probability distribution P is

$$\mathbb{E}_P[f] = \sum_a P(a) f(a)$$

$$\mathbb{E}_p[f] = \int_x p(x) f(x)\, dx$$

In particular if $f(x) = x$, this is the mean of the random variable x.

The variance of f is $\mathrm{Var}(f) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$. The standard deviation is $std(f) = \sqrt{\mathrm{Var}(f)}$.

The covariance between x, y is $\mathrm{Cov}(x, y) = \mathbb{E}_{x,y}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] = \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]$.

x, y can be vectors. $\mathbb{E}[x]$ is the mean vector. $\mathrm{Cov}(x, y)$ is the covariance matrix with i, j-th entry being $\mathrm{Cov}(x_i, y_j)$.

2 Distributions

Uniform distribution with K outcomes $P(A = a_i) = 1/K$, $i = 1, \ldots, K$.

Bernoulli distribution on binary variable $x \in \{0, 1\}$: $P(x|\mu) = \mu^x (1 - \mu)^{(1-x)}$. Mean $\mathbb{E}[x] = \mu$, variance $\mathrm{Var}(x) = \mu(1 - \mu)$.

Binomial distribution: the probability of observing m heads in N trials. $P(m|N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$, with $\binom{N}{m} = \frac{N!}{(N-m)!m!}$. $\mathbb{E}[m] = N\mu$, $\mathrm{Var}(m) = N\mu(1-\mu)$.

Multinomial distribution for K-sided die with probability $\mu = (\mu_1, \ldots, \mu_K)$ which sum to 1, N throws, counts $m_1, \ldots, m_K$:

$$P(m_1, \ldots, m_K | \mu, K) = \binom{N}{m_1 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k}.$$

Dirichlet distribution (the dice factory) on μ, with 'hyper'-parameters $\alpha > 0$:

$$p(\mu|\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}.$$

The gamma function is a generalization of factorial, with the property $\Gamma(x+1) = x\Gamma(x)$ and $\Gamma(1) = 1$. Dirichlet is the conjugate prior for multinomial.

Gaussian (Normal) distributions
univariate: $p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. Mean μ, variance σ^2.
multivariate: $p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$, where x, μ are D-dimensional vectors, Σ is a $D \times D$ covariance matrix.

3 Linear Algebra

Scalar (1×1), vector (default column vector, $n \times 1$), matrix ($n \times m$). Matrix transpose $\left(A^\top\right)_{ij} = A_{ji}$.

A $n \times m$ matrix A times a $m \times p$ matrix B is a $n \times p$ matrix C, with $C_{ij} = \sum_{k=1}^{m} A_{ik} B_{kj}$. Check dimensions.

$(AB)C = A(BC), A(B+C) = AB + AC, (A+B)C = AC + BC, (A+B)^\top = A^\top + B^\top, (AB)^\top = B^\top A^\top$. Note in general $AB \neq BA$.

The following is specific to square matrices.

Diagonal matrix: $A_{ij} = 0, \forall i \neq j$. Identity matrix I is diagonal with $I_{ii} = 1, \forall i$. $AI = IA = A$ for all square A.

Some square matrices have inverses: $AA^{-1} = A^{-1}A = I$. $(AB)^{-1} = B^{-1}A^{-1}$. $(A^\top)^{-1} = (A^{-1})^\top$.

The trace is the sum of diagonal elements (or eigenvalues) $\mathrm{Tr}(A) = \sum_i A_{ii}$.

The determinant $|A|$ is the product of eigenvalues. $|AB| = |A||B|$, $|a| = a$, $|aA| = a^n|A|$, $|A^{-1}| = 1/|A|$. A matrix A is invertible iff $|A| \neq 0$.

If $|A| = 0$ for a $n \times n$ square matrix A, A is said to be singular. This means at least one column is linearly dependent on (i.e., a linear combination of) other columns (same for rows). Once all such linearly dependent columns and rows are removed, A is reduced to a smaller $r \times r$ matrix, and r is called the rank of A.

A $m \times m$ matrix A has m eigenvalues λ_i and eigenvectors (up to scaling) u_i s.t. $Au_i = \lambda_i u_i$. In general λ's are complex numbers. If A is real and symmetric, λ's are real numbers, and u's are orthogonal. The u's can be scaled to orthonormal, i.e., length one, so that $u_i^\top u_j = I_{ij}$. The spectral decomposition is $A = \sum_i \lambda_i u_i u_i^\top$. For invertible A, $A^{-1} = \sum_i \frac{1}{\lambda_i} u_i u_i^\top$. This shows why the determinant must be non-zero.

A real symmetric matrix A is positive semi-definite, if its eigenvalues $\lambda_i \geq 0,\ \forall i$. Equivalently, $\forall x \in \mathbb{R}^n, x^\top A x \geq 0$. It is strictly positive definite if $\lambda_i > 0,\ \forall i$.

A positive semi-definite matrix has rank r equal to the number of positive eigenvalues. The remaining $n - r$ eigenvalues are zero.

For vector $x \in \mathbb{R}^n$, we have
0-norm: $\|x\|_0 =$ count of nonzero elements
1-norm: $\|x\|_1 = \sum_{i=1}^n |x_i|$
2-norm (the Euclidean norm, or just 'the norm', length: $\|x\|_2 = \left(\sum_{i=1}^n x_i^2\right)^{1/2}$
∞-norm: $\|x\|_\infty = \max_{i=1}^n |x_i|$

4 Calculus

The derivative (slope of tangent line) is $f'(x) = \frac{df}{dx} = \lim_{\delta \to 0} \frac{f(x+\delta) - f(x)}{\delta}$.

The second derivative (curvature) is $f''(x) = \frac{d^2 f}{dx^2} = \frac{df'}{dx}$.

Often used ones: $c' = 0$, $(cx)' = c$, $(cx^a)' = cax^{a-1}$, $(\log x)' = 1/x$, $(e^x)' = e^x$, $(f(x) + g(x))' = f'(x) + g'(x)$, $(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$.

The chain rule: $\frac{df(y)}{dx} = \frac{df(y)}{dy} \frac{y}{dx}$.

For multivariate function $f(x_1, \ldots, x_n)$, the partial derivative w.r.t. x_i is

$$\frac{\partial f}{\partial x_i} = \lim_{\delta \to 0} \frac{f(x_1, \ldots, x_i + \delta, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_i, x_{i+1}, \ldots, x_n)}{\delta}.$$

The gradient at $x = (x_1, \ldots, x_n)$ is

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}.$$

The gradient is a vector in the same space as x. It points to 'higher ground' in terms of f value.

The second derivatives form a $n \times n$ Hessian matrix

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

.

Unconstrained optimality conditions:

- necessary conditions: local minima have $\nabla f(x) = 0$ and $\nabla^2 f(x)$ positive semidefinite.
- sufficient conditions: Any point x at which $\nabla f(x) = 0$ and $\nabla^2 f(x)$ positive definite is a local minimum.

A function f is convex, if $\forall x, y, \lambda \in [0, 1]$, $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$. Common convex functions: c, cx, $(x-c)^n$ if n is an even integer, $|x|$, $1/x$, e^x. When the second derivative exists, it is non-negative (positive semi-definite Hessian).

A function f is concave, if $\forall x, y, \lambda \in [0, 1]$, $f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda) f(y)$. Common concave functions: c, cx, $-(x-c)^n$ if n is an even integer, $\log x$. When the second derivative exists, it is non-positive (negative semi-definite Hessian).

If f is convex and differentiable, $\nabla f(x) = 0$ iff x is a global minimum.