Probability and linear algebra review

Linear algebra:

Matrix: arrangement of numbers with some useful properties
useful to represent sets of linear eqns.

$$A \in \mathbb{R}^{m\times n} \equiv \begin{array}{c} m \\ \text{rows} \end{array} \left\{ \underbrace{\begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1} & \cdots & A_{mn} \end{bmatrix}}_{n \text{ cols}} \right.$$

Vector: matrix with either 1 column or row.

$$x \in \mathbb{R}^n \equiv \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Summation and subtraction: element wise.
Element wise multiplication: Hadamard product $(A \circ B)$.
Transpose $(A^T)_{ij} = A_{ji}$.
Matrix multiplication: $C = AB$ where $C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$

$$A \in \mathbb{R}^{m\times n} \qquad B \in \mathbb{R}^{n\times p} \qquad \begin{bmatrix} 0 & \\ & \end{bmatrix} = \begin{bmatrix} \longrightarrow \\ \; \end{bmatrix} \begin{bmatrix} \downarrow & \end{bmatrix}$$

Vector-vector product:

$$x^T y = \sum_{i=1} x_i y_i \quad : \text{scalar product, inner product}$$

$$x y^T = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} [y_1 \cdots y_n] = \begin{bmatrix} x_1 y_1 & \cdots & x_1 y_n \\ \vdots & \ddots & \vdots \\ x_m y_1 & \cdots & x_m y_n \end{bmatrix}$$

: outer product.

Useful to do things like:

$$A = \begin{bmatrix} | & | & & | \\ x & x & \cdots & x \\ | & | & & | \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} = x 1^T$$

with $1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$

Matrix vector products: $y = Ax = \begin{bmatrix} A_{1*} \\ A_{2*} \\ \vdots \\ A_{m*} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} A_{1*} x_1 \\ A_{2*} x_2 \\ \vdots \\ A_{m*} x_n \end{bmatrix}$

linear combination: $y = \begin{bmatrix} A_{*1} & A_{*2} & \cdots & A_{*n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$

$= \begin{bmatrix} A_{*1} \end{bmatrix} x_1 + \cdots + \begin{bmatrix} A_{*n} \end{bmatrix} x_n$

Properties: $(AB)C = A(BC)$

$A(B+C) = AB + AC$

not commutative

$AI = IA = A$ with $I = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix} = \text{diag}(1, \cdots, 1)$

$(AB)^T = B^T A^T \rightarrow (A_1 A_2 \cdots A_k)^T = A_k^T A_{k-1}^T \cdots A_2^T A_1^T$

$(A+B)^T = A^T + B^T$

$B = A^{-1}$ iff $AA^{-1} = A^{-1}A = I$: A square

A non-invertible $\Leftrightarrow$ A singular $\Leftrightarrow$ $\det(A) = 0$

$(A^{-1})^T = (A^T)^{-1}$, $(AB)^{-1} = B^{-1}A^{-1}$

pseudo-inverses: $\underset{\text{left}}{(A^TA)^{-1}A^T}$ or $\underset{\text{right}}{A^T(AA^T)}$

Symmetric matrices : $A^T = A$.
Orthogonal matrices : $A^T = A^{-1}$
Trace : $tr(A) = \sum_{i=1}^{n} A_{ii}$
with $tr A = tr A^T$
$tr(A+B) = tr A + tr B$
$tr(tA) = t\, tr A$
$tr AB = tr BA$ iff AB is square

Norm : informally known as length or magnitude of a vector.

$l_2 : \|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$
$l_1 : \|x\|_1 = \sum_{i=1}^{n} |x_i|$

$\infty : \|x\|_\infty = \max_i |x_i|$.
$l_p : \|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$

Frobenius $\|A\|_F = \sqrt{tr(A^T A)}$
Normalized : $\|x\|_2 = 1$
Linear independence : $\{x_1, \ldots, x_n\} \subset \mathbb{R}^m$
no vector can be expressed as a linear combination of the remaining ones
linearly dependent. $\alpha_1, \ldots, \alpha_{n-1}$ exist st. $x_n = \sum_{i=1}^{n-1} \alpha_i \bar{x}_i$

column rank : size of largest subset of columns that form a linearly independent set
row rank : similar but for rows
row rank = column rank = rank (A)
rank$(A) \le \min(m, n)$. Full rank $\leftrightarrow$ equality.

$\text{span}(\{x_1, \ldots, x_n\}) = \{v : v = \sum_{i=1}^{n} \alpha_i x_i \quad \alpha_i \in \mathbb{R}\}$

if $\{x_1, \ldots, x_n\}$ are linearly independent, span $\to \mathbb{R}^n$.

range$(A) =$ span$(\{A_{*1}, A_{*2}, \ldots, A_{*n}\})$. Nullspace $N(A) = \{x \in \mathbb{R}^n : Ax = 0\}$

Determinant:

volume of parallelotope : geometric interpretation.

$\det(I) = 1$

$|A| = |A^T|$

$|AB| = |A||B|$

$|A| = 0$ iff. A is singular.

$|A^{-1}| = 1/|A|$

Quadratic form and positive definiteness

A: square $\in \mathbb{R}^{n \times n}$, $x \in \mathbb{R}^n$

$$x^T A x = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

$$x^T A x = (x^T A x)^T = x^T A^T x =$$

$$\frac{1}{2} x^T A x + \frac{1}{2} x^T A^T x = x^T A x$$

$$\frac{1}{2} x^T (A + A^T) x = x^T A x$$

$$x^T \left(\frac{1}{2}A + \frac{1}{2}A^T\right) x = x^T A x$$

PD iff A is symmetric and $x^T A x > 0$ for any x.

ND etc...

Diagonalization and eigenvectors

$$A\underset{\text{eigenvalue}}{\lambda}\overset{\text{eigenvector}}{x}$$

$$Ax = \lambda x$$

eigenvector — eigenvalue

$$|A| = \prod_{i=1}^{n} \lambda_i$$

Diagonalize : $A = Q \Lambda Q^{-1}$

Symmetric : $A = Q \Lambda Q^{T}$

Gradients:

Let $f(x) = \sum_{i=1}^{n} b_i x_i = b^T x$

$$\frac{\partial f(x)}{\partial x} = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} . \quad \frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} b_i x_i = b_i$$

$$\therefore \frac{\partial f(x)}{\partial x} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} = b \quad \text{or} \quad \nabla_x b^T x = b.$$

Let $f(x) = x^T A x$.

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} x_i x_j = \frac{\partial}{\partial x_k}\Big[\sum_{i\neq k}\sum_{j \neq k} A_{ij} x_i x_j + \sum_{i \neq k} A_{ik} x_i x_k + \sum_{j\neq k} A_{kj} x_k x_j + A_{kk} x_k^2$$

$$= \sum_{i\neq k} A_{ik} x_i + \sum_{j\neq k} A_{kj} x_j + 2A_{kk} x_k$$

$$= \sum_{i=1}^{n} A_{ik} x_i + \sum_{j=1}^{n} A_{kj} x_j \quad \therefore \nabla_x f(x) = A^T x + A x$$

$$= 2Ax$$

:

$$J_x y = \frac{\partial y}{\partial x^T} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

$$\nabla_x y = \frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix}$$

$$\frac{\partial A x}{\partial x^T} = A \quad ; \quad \frac{\partial x^T x}{\partial x^T} = 2x$$

$\rightarrow$

Probability review:

Frequentist notion: # of occurences / # trials

Laplace view

Bayesian notion: uncertainty $\rightarrow$ related to information

Good for rare events. Eg: prob. of Red Sox winning World Series this year.

Kolmogorov axioms:

Event A with some outcome

(i) $P(A) \geq 0$

(ii) $P(\Omega) = 1$

(iii) if $A \cap B = \emptyset$, $P(A \cup B) = P(A) + P(B)$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

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

in other words

$$p(A|B) = \frac{\overbrace{p(A \cap B)}^{\text{joint}}}{\underbrace{p(B)}_{\text{marginal}}}$$

its a probability

$$p(A) = \sum_b p(A,B) = \sum_b p(A|B=b)\,p(B=b)$$

$$p(X_1 X_2 \cdots X_D) = p(X_1)\,p(X_2|X_1)\,p(X_3|X_2,X_1) \cdots p(X_D|X_1,\ldots,X_{D-1})$$

Discrete random variables:

$p(X=x)$, ~~XXXR~~ $x \in \mathcal{X} \subset \mathbb{R}$

Bayes Rule: $p(X=x \mid X=y) = \dfrac{p(X=x, Y=y)}{p(Y=y)}$

$$= \frac{p(X=x)\,p(Y=y|X=x)}{\sum_{x'} p(X=x')\,p(Y=y|X=x')}$$

Eg: medical diagnosis

Woman in 40s. Mammogram. If test is positive, do you have cancer? Let's see if it's reliable

sensitivity of 80% i.e. if you have cancer, test will be positive 8

$p(x=1 \mid y=1) = 0.8$; $x=1$: test +.
$y=1$: cancer

People think that they have 80% chances of having cancer!

prior $p(y=1) = 0.004$: prob. that population has BC.

$p(x=1 \mid y=0) = 0.1$; 10% chance to give a false positive

$p(y=1 \mid x=1) = \dfrac{p(x=1|y=1)\,p(y=1)}{\quad}$ $= 0.031$

Independence : $p(x, y) = p(x)\, p(y) \Leftrightarrow X \perp Y$

Conditionally independent $X \perp Y \mid Z \Leftrightarrow p(x, y \mid z) = p(x \mid z)\, p(y \mid z)$

Continuous random variables

pdf - $\cancel{f(x)}$

cumulative dist. function $P(a < X \leq b) = \int_a^b f(x)\, dx$

Expected value (mean)

$$E(X) = \int_{\mathcal{X}} x\, f(x)\, dx$$

$$= \sum_{x \in \mathcal{X}} x\, p(X = x)$$

Variance (spread) : $var(X) = E[(X - \mu_X)^2]$

$$= E[X^2] - E[X]^2$$

Covariance : $cov(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$

Gaussian distribution $\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$

$$\mathcal{N}(X \mid \mu, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

$$Cov(X) = E[(x - E(x))(x - E(x))^T]$$

Linear transformations :

X : RVect. distribution of $y = Ax + b$

$$E[y] = E[Ax + b] = A\mu_x + b$$

$$cov(Y) = cov(Ax + b) = A \Sigma_x A^T$$