Discussion 1
Prof. J. Kao
TAs: T. Monsoor, T. Wang, P. Lu, Y. Li

1 Probability basics

In statistical signal processing, manipulation of probability expressions is very important. The two tools from probability theory that we will be using frequently to manipulate the expressions are:

- Law of total probability
- Probability chain rule

1.1 Law of total probability

If $A_1, A_2, \ldots, A_n$ forms a partition of the sample space S, then the probability of an event B is given by

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i)$$

Using the definition of conditional probability, we can rewrite the above expression as

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Thus, the probability of the event B is the sum of the conditional probabilities $P(B|A_i)$ weighted by the probability $P(A_i)$.

1.2 Probability chain rule

For any events A and B, we have

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

More generally, for any events $A_1, A_2, \ldots, A_n$ we get

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = \prod_{k=1}^{n} P(A_k | \cap_{j=1}^{k-1} A_j)$$

You can use induction to show the general version of the probability chain rule.

1.3 Practice problem on probability basics

A computer manufacturer uses chips from three sources. Chips from sources A, B, and C are defective with probabilities 0.005, 0.001, and 0.010, respectively. If a randomly selected chip is found to be defective, find:

1. The probability that the manufacturer was A
2. The probability that the manufacturer was C

Assume that the proportions of chips from A, B, and C are 0.5, 0.1, and 0.4, respectively.

2 Linear Algebra basics

Proficiency in linear algebra is required to have a solid foundation in machine learning. The two linear algebra concepts that we will be using repeatedly in this course are:

- Decomposition of matrices ← Touch in todays discussion
- Derivatives of vectors and matrices ← Focus of Disc. 2

In this discussion we will focus on the decomposition of matrices, namely:

- Eigendecomposition
- Singular value decomposition

Δ: Diagonal matrix with e-values as it's elements

2.1 Eigendecomposition

Q: orthogonal matrix with colums as eigenvectors.

Every real symmetric matrix, $A \in \mathbf{R}^{n\times n}$, can be factored as

$$A = Q\Delta Q^T$$

where $\Delta = diag(\lambda_1, \lambda_2, \cdots, \lambda_n) \in \mathbf{R}^{n\times n}$. More specifically we have,

- The set $\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$ are called the eigenvalues of A and can be found by solving the equation

$$\boxed{det(A - \lambda I) = 0}$$

← Solving chr. eqn

- The columns of Q are an orthonormal set of n eigenvectors and can be found by solving the equation

$$\boxed{Av_i = \lambda_i v_i}$$

$QQ^T = Q^TQ = I$ ← Orthogonal matrix

rank: # of linearly independent columns/rows

2.2 Singular value decomposition

Let $A \in \mathbf{R}^{m\times n}$ be a matrix with rank r, then there exists orthogonal matrices $U \in \mathbf{R}^{m\times m}$ and $V \in \mathbf{R}^{n\times n}$ such that

$\Sigma \in \mathbb{R}^{m\times n}$

$m\times m$ $\quad m\times n$ $\quad n\times n$

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

$$\Sigma = \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix}$$

where $\Sigma = \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix}$, $S = diag(\sigma_1, \sigma_2, \cdots, \sigma_r) \in \mathbf{R}^{r\times r}$, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. More specifically, we have

- The set $\{\sigma_1, \sigma_2, \cdots, \sigma_r\}$ are called the non-zero singular values of A and can be calculated as

$$\sigma_i(A) = \lambda_i^{\frac{1}{2}}(A^TA) = \lambda_i^{\frac{1}{2}}(AA^T)$$

- The columns of U are called the left singular vectors of A (and are the orthonormal eigenvectors of AA^T).
- The columns of V are called the right singular vectors of A (and are the orthonormal eigenvectors of A^TA).

2.3 Practice problem on linear algebra basics

1. Show the following properties for matrices

   (a) If $b^TAb > 0$ for all $b \in \mathbf{R}^n$, then all eigenvalues of A are positive.

   (b) If $A \in \mathbf{R}^{n\times n}$ is an orthogonal matrix, then all eigenvalues of A have norm 1.

   (c) If $A \in \mathbf{R}^{m\times n}$ is a matrix with rank r, then

$$\sigma_i(A) = \lambda_i^{\frac{1}{2}}(AA^T)$$

$\|x\|_2^2 = x^Tx$

$b^TAb > 0$ for all $b \in \mathbb{R}^n$

a) Let λ_i be an e-value of A with corresponding e-vector v_i. Then,

$$Av_i = \lambda_i v_i$$

$$v_i^T A v_i = v_i^T \lambda_i v_i$$

$$\Rightarrow v_i^T A v_i = \lambda_i v_i^T v_i$$

$$\Rightarrow v_i^T A v_i = \lambda_i \|v_i\|_2^2$$

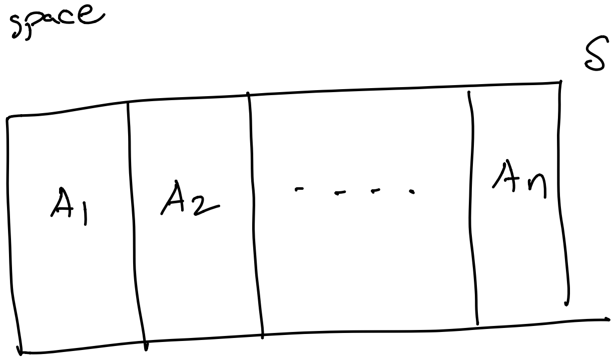
$$\Rightarrow \lambda_i \|v_i\|_2^2 > 0 \Rightarrow \boxed{\lambda_i > 0}$$

Since $b^TAb > 0$ for all $b \in \mathbb{R}^n$

Then, $v_i^TAv_i > 0$

Law of total probability:

Suppose $A_1, A_2, \ldots A_n$ forms a partition of my sample space



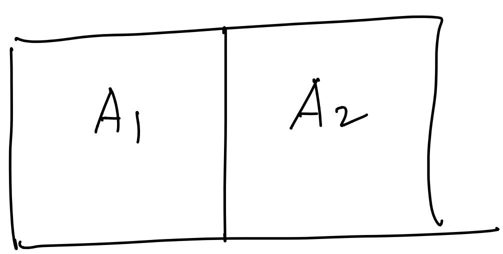
- $A_1 \cup A_2 \cup \ldots \cup A_n = S$
- $A_i \cap A_j = \emptyset$ for $i \neq j$

Suppose I flip a coin

A_1: outcome of my coin flip is heads

A_2: outcome of my coin flip is tails

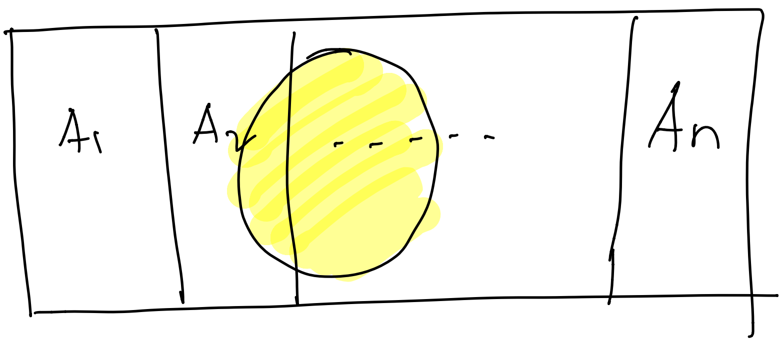
A_2: outcome of my coin



$\Rightarrow A_1 \cup A_2 = S$

$\Rightarrow A_1 \cap A_2 = \emptyset$

then
$$\boxed{P(B) = \sum_{i=1}^{n} P(B \cap A_i)}$$



$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \dots + P(B \cap A_n)$$

Using definition of cond. prob,

$$P(B \cap A_i) = P(B \mid A_i)\, P(A_i)$$

Then,

$$\boxed{P(B) = P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + \dots + P(B \mid A_n)P(A_n)}$$

Prob. chain rule:

Suppose we have n events

$$A_1, A_2, A_3, \dots\dots, A_{n-1}, A_n$$

Then,

$$P(A_1 \cap A_2 \cap \dots \cap A_{n-1} \cap A_n)$$

$$= \underset{\varphi}{P(A_1)}\, \underset{\varphi}{P(A_2 \mid A_1)}\, \underset{\varphi}{P(A_3 \mid A_1, A_2)} \dots P(A_n \mid A_1, A_2, \dots A_{n-1})$$

Example

$$P(A_1, A_2, A_4)$$

$$= P(A_1)\, P(A_2 \mid A_1)\, P(A_4 \mid A_2, A_1)$$

$$= P(A_4)\, P(A_2 \mid A_4)\, P(A_1 \mid A_2, A_4)$$

$$= P(A_2)\, P(A_4 \mid A_2)\, P(A \mid A_2, A_4)$$

$$=$$

Let's define the following events:

Z_D: A randomly selected chip is defective

Z_A: A randomly selected chip is manufactured by A

Z_B: A randomly selected chip is manufactured by B

Z_C: A randomly selected chip is manufactured by C

Now we want to compute

$P(Z_A \mid Z_D)$, $P(Z_C \mid Z_D)$

Given:

$P(Z_D \mid Z_A) = 0.005$

$P(Z_D \mid Z_B) = 0.001$

$P(Z_D \mid Z_C) = 0.01$

$$\boxed{P(Z_A \mid Z_D) = \frac{P(Z_D \mid Z_A)\, P(Z_A)}{P(Z_D)}}$$

$P(Z_A, Z_D) = P(Z_D)\, P(Z_A \mid Z_D)$

$= P(Z_A)\, P(Z_D \mid Z_A)$

Need to figure this out

Since Z_A, Z_B and Z_C form a partition of sample space, so by law of total prob.

$$P(Z_D) = P(Z_D \mid Z_A)\, P(Z_A) + P(Z_D \mid Z_B)\, P(Z_B) + P(Z_D \mid Z_C)\, P(Z_C)$$

$$= 0.0066$$

$$\therefore P(Z_A \mid Z_D) = \frac{0.005 \times 0.5}{0.0066} = 0.3788$$

b) Let λ_i be an e-value of A with corresponding e-vector v_i. Then

$$A v_i = \lambda_i v_i$$

$$A^T A v_i = \lambda_i A^T v_i$$

$$\Rightarrow \quad v_i = \lambda_i A^T v_i$$

$$\|v_i\|_2^2 = \|\lambda_i A^T v_i\|_2^2$$

$$= |\lambda_i|^2 \|A^T v_i\|_2^2$$

$$\|A^T v_i\|_2^2 = (A^T v_i)^T (A^T v_i)$$

$$= v_i^T A A^T v_i$$

$$= v_i^T v_i = \|v_i\|_2^2$$

$$\therefore \quad \|v_i\|_2^2 = |\lambda_i|^2 \|v_i\|_2^2$$

$\Rightarrow \quad |\lambda_i| = 1$. Hence all e-values have magnitude 1.

(c) $A \in \mathbb{R}^{m \times n}$ with rank r

σ_i : i^{th} singular value of A

Then,

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

Now,

$$AA^T = (U\Sigma V^T)(U\Sigma V^T)^T$$

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

$$AA^T = U \boxed{\Sigma \Sigma^T} U^T$$

$$\Sigma = \begin{bmatrix} \sigma_1 & & & & & & \mathbf{0} \\ & \sigma_2 & & & & & \\ & & \ddots & & & & \\ & & & \sigma_r & & & \\ & & & & 0 & & \\ & \mathbf{0} & & & & \ddots & \\ & & & & & & 0 \end{bmatrix}$$

$$\Sigma^T = \begin{bmatrix} \sigma_1 & & & & & & \mathbf{0} \\ & \sigma_2 & & & & & \\ & & \ddots & & & & \\ & & & \sigma_r & & & \\ & & & & 0 & & \\ & \mathbf{0} & & & & \ddots & \\ & & & & & & 0 \end{bmatrix}$$

$$\Sigma \Sigma^T = \begin{bmatrix} \sigma_1^2 & & & & & & \mathbf{0} \\ & \sigma_2^2 & & & & & \\ & & \ddots & & & & \\ & & & \sigma_r^2 & & & \\ & & & & 0 & & \\ & \mathbf{0} & & & & \ddots & \\ & & & & & & 0 \end{bmatrix}$$

Since $\Sigma\Sigma^T$ is a diagonal matrix and U is an orthogonal matrix so eigendecomposition of AA^T is

$$\boxed{AA^T = \overset{\text{orth.}}{U}\,\boxed{\Sigma\Sigma^T}\,U^T}$$

$\sigma_1^2, \sigma_2^2, \ldots \sigma_r^2, 0, \ldots 0$

Then,

$$\lambda_i(AA^T) = \sigma_i^2(A)$$

$$\Rightarrow \boxed{\sigma_i(A) = \lambda_i^{1/2}(AA^T)}$$

$$b^T(AA^T)b$$

$$= b^TAA^Tb$$

$$= (A^Tb)^T(A^Tb)$$

$$= \|A^Tb\|_2^2 \geq 0 \quad \text{PSD}$$

$A \in \mathbb{R}^{m \times n}$

$\boxed{\text{rank} \le \min(m, n)}$

If $\text{rank} = \min(m, n)$

then A is full rank.