RECITATION — 02/17/2010

LINEAR ALGEBRA TUTORIAL

→ Why bother? Sophisticated machine probabilistic models have a large number of parameters and define distributions on multi-dimensional random variables. Keeping track of these quantities is tedious. We want operators which can conviniently operate on collections of r.v.s.

→ Notation for this talk:

$x_j \in \mathbb{R}^1$ a scalar

$X \in \mathbb{R}^{N \times 1}$ $\begin{bmatrix} \\ \\ \\ \end{bmatrix}$, column vector

$X^T \in \mathbb{R}^{1 \times N}$ (transpose) $[\quad\quad]$; row vector

$X \in \mathbb{R}^{M \times N}$ is a matrix.

$$\begin{bmatrix} — a_1^T — \\ \vdots \\ — a_M^T — \end{bmatrix} \text{ or equivalently } \begin{bmatrix} | & & | \\ b_1 & \cdots & b_N \\ | & & | \end{bmatrix}$$

→ Basic operations:

a) Vector - Vector operations:

$Y, X \in \mathbb{R}^{N \times 1}$

$$X^T Y = \langle X, Y \rangle = \begin{bmatrix} x_1 & \cdots & x_N \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \sum_i x_i y_i \in \mathbb{R}^1$$

$$XY^T = \begin{bmatrix} x_1 y_1 & \cdots & x_1 y_n \\ & & \\ x_n y_1 & & x_n y_n \end{bmatrix}$$

rank 1 matrix.

Now remember covariance between two random variable:

$$\text{cov}[x_1, x_2] = E[(x_1 - \mu_{x_1})(x_2 - \mu_{x_2})]$$

$$= E[x_1 x_2] - E[x_1] E[x_2]$$

measures the degree to which x_1 & x_2 are related.

$$\text{cov}[x, x] = E[x^2] - E[x]^2 = \text{var}(X)$$

For vectors:-

$X \in R^{N \times 4}$ we have $[x_1, x_2 \cdots x_D]^T$ we have.

$$\text{Cov}[X] = E[(X - \mu_X)(X - \mu_X)^T]$$

$$= E[XX^T] - E[$$

$$E[(X - \mu_X)(X^T - \mu_X^T)]$$

$$= E[XX^T - \mu_X X^T - X\mu_X^T + \mu_X \mu_X^T]$$

$$= E[xx^T] - \mu_x E[x^T] - E[x]\mu_x^T + \mu_x \mu_x^T$$

$$= E[xx^T] - 2\mu_x \mu_x^T + \mu_x \mu_x^T$$

$$= E[xx^T] - \mu_x \mu_x^T$$

$$= \begin{bmatrix} E[x_1^2] & E[x_1 x_2] & \cdots & E[x_1 x_n] \\ E[x_2 x_1] & & & \\ \vdots & & & \\ E[x_n x_1] & & & E[x_n^2] \end{bmatrix} - \begin{bmatrix} E[x_1]^2 & E[x_1]E[x_2] & \cdots & E[x_1]E[x_n] \\ E[x_2]E[x_1] & & & \\ \vdots & & & \\ E[x_n]E[x_1] & & & E[x_n]^2 \end{bmatrix}$$

$$= \begin{bmatrix} E[x_1^2] - E[x_1]^2 & E[x_1 x_2] - E[x_1]E[x_2] & \cdots & \\ & & & \\ & & & \end{bmatrix}$$

$$= \begin{bmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1, x_2) & \cdots & \mathrm{cov}(x_1, x_n) \\ \mathrm{cov}(x_n, x_1) & & & \mathrm{var}(x_n) \end{bmatrix}$$

The values in the covariance matrix

⇒ look at correlation matrices.

b) Matrix Vector products:

Two equivalent interpretations, but it might help to think of the product in one or the other way depending on the problem

$$\underset{A}{\begin{bmatrix} | & & | \\ a_1 & \cdots & a_n \\ | & & | \end{bmatrix}} \underset{x}{\begin{bmatrix} x_1 \\ \\ x_n \end{bmatrix}} = \underset{b}{} \begin{bmatrix} | \\ a_1 \\ | \end{bmatrix} x_1 + \begin{bmatrix} | \\ a_2 \\ | \end{bmatrix} x_2 + \cdots \begin{bmatrix} | \\ a_n \\ | \end{bmatrix} x_n$$

1)

2)
$$\begin{bmatrix} - a_1^T - \\ \\ - a_M^T - \end{bmatrix} \begin{bmatrix} \\ x \\ \\ \end{bmatrix} = \begin{bmatrix} a_1^T x \\ \vdots \\ a_M^T x \end{bmatrix}$$

interpretation 1) ⇒ x acts on A ; A is the data & x is the operator.

for instance consider :

two random vectors a_1 & a_2 and want a convex combination of the two.

$$\begin{bmatrix} | & | \\ a_1 & a_2 \\ | & | \end{bmatrix} \begin{bmatrix} P \\ \\ 1-P \end{bmatrix} = \begin{bmatrix} | \\ a_1 \\ | \end{bmatrix} P + \begin{bmatrix} | \\ a_2 \\ | \end{bmatrix} (1-P)$$

interpretation of 2): A acts on the vector x ; x → is the data & A is the operator.

$$\begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix} \begin{bmatrix} \\ x \\ \\ \end{bmatrix} = \begin{bmatrix} x_1 \\ \\ 10x_2 \end{bmatrix}$$

any transformation matrix fits the second interpretation

c) Matrix-Matrix products:

follow naturally from matrix vector products

$AB \neq BA$

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

$A(BC) = (AB)C$

d) Orthogonal Matrices:

A square matrix $U \in \mathbb{R}^{N \times N}$ is orthogonal if all its columns/rows are orthonormal.

Orthonormal vectors $x^T y = 0$ & $\|x\|_2 = 1 = \|y\|_2$
where $\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$

Orthogonal matrices: $U^T U = U U^T = I$
$\Rightarrow U^{-1} = U^T$

e) Eigenvalues and Eigenvectors.

Given $A \in \mathbb{R}^{N \times N}$, a vector $v \in \mathbb{R}^N$ and $\lambda \in \mathbb{R}^1$ are the eigenvectors and eigenvalues

$$Av = \lambda v \; ; \; v \neq 0$$

Transforming a vector v by A, results in a scaled vector with unchanged orientation.

Since A can only scale the vector it is immediately apparent that A has to be a square matrix.

If A is real and symmetric all it's eigenvalues are real and it's eigen vectors are orthonormal.

$$Au = \lambda u$$

we can collect all the eigenvectors and values

$$AU = U\Lambda$$

where $U = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_N \\ | & & | \end{bmatrix}$ $\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{bmatrix}$

now for the symm

$$A = U\Lambda U^{-1} = U\Lambda U^T = \sum_{i=1}^{N} \lambda_i u_i u_i^T$$

Approximating Matrix A

Note that $A = \sum_{i=1}^{N} \lambda_i u_i u_i^T$; $\lambda_1 \geq \lambda_2 \geq \lambda_3 \cdots$

Often we are interested in a lower dimensional approximation of A; $\tilde{A}$

$$\tilde{A} = \sum_{i=1}^{D} \lambda_i u_i u_i^T \quad ; \quad D < N \; ;$$

$$= \begin{bmatrix} | & & | \\ u_1 & \cdots & u_D \\ | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_D \end{bmatrix} \begin{bmatrix} - & u_1^T & - \\ & \vdots & \\ - & u_D^T & - \end{bmatrix}.$$

related to the ideas of PCA; can be shown to be the "best" rank D approximation according to ~~some~~ a objective.

Positive Definite matrices

Why? Consider the multivariate Gaussian: $x \in \mathbb{R}^D$

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

$$\mu = \vec{0}$$

$$\mathcal{N}(x|0,\Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}x^T\Sigma^{-1}x\right\}.$$

Consider what happens when $x^T\Sigma^{-1}x < 0$
we can no longer appropriately normalize the distribution

A symmetric matrix A is called positive definite iff for all $x \neq 0$, $x^TAx > 0$. Thus we need Σ^{-1} and thus Σ to be positive definite for a proper normal distribution.

Proof

$\Rightarrow$ A matrix is positive ~~semi~~ definite iff all its eigenvalues are > 0

Show $\forall i \lambda_i > 0 \Rightarrow x^TAx > 0$.

Proof: $x^TAx = x^TU\Lambda U^Tx = y^T\Lambda y = \sum_{i=1}^{N}\lambda_i y_i^T y_i$

$= \sum_{i=1}^{N}\lambda_i y_i^2 \Rightarrow$ iff $\forall \lambda_i > 0 \Rightarrow x^TAx > 0$.

$\Rightarrow$ A, symmetric matrix is P.D. iff all its eigenvalues are > 0.

Proof: $\Rightarrow$ $x^TAx > 0 \Rightarrow \lambda_x > 0$

For any eigenvector $x \neq 0$ we know $x^TAx > 0$

$\Rightarrow x^T \lambda_x x > 0$

$\lambda_x (x^Tx) > 0$

$\therefore \lambda_x > 0$

$\Leftarrow$ if all $\lambda_i > 0 \Rightarrow x^TAx > 0$

for some $x \neq 0$ $\Rightarrow x^TAx = x^TU\Lambda U^Tx = y^T\Lambda y$ with $y = x^TU$

$$y^T\Lambda y = \sum_{i=1}^{N} \lambda_i y_i^T y_i = \sum_{i=1}^{N} \lambda_i y_i^2$$

Now if all $\lambda_i > 0 \Rightarrow \sum_{i=1}^{N} \lambda_i y_i^2 > 0$ since $y_i > 0$

$\Rightarrow x^TAx > 0$

$\Rightarrow$ A symmetric matrix with all $\lambda_i \geq 0$ is positive semi-definite and one with all $\lambda_i < 0$ is negative definite. Everything else is indefinite.

Given a real symmetric matrix you can determine it's definiteness by

a) Calculating its eigenvalues

b) Sometimes you can tell by looking at the matrix

→ A sufficient but not necessary condition for a matrix to be positive definite is diagonal dominance

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \quad \text{for all } i$$

$$\begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Other properties of eigenvalues

1) $|A| = \prod_{i=1}^{N} \lambda_i \quad \Rightarrow \quad \log|A| = \sum_{i=1}^{N} \log \lambda_i$

2) $\text{tr}(A) = \sum_{i=1}^{N} \lambda_i$

5 Matrix Inverses.

Consider a trivariate Gaussian distribution

$$\mathcal{N}\left(\begin{bmatrix} x_a \\ x_b \\ x_c \end{bmatrix} \middle| \begin{bmatrix} \mu_a \\ \mu_b \\ \mu_c \end{bmatrix}, \Sigma\right) = p\left(\begin{bmatrix} x_a \\ x_b \\ x_c \end{bmatrix} \middle| \begin{bmatrix} \mu_a \\ \mu_b \\ \mu_c \end{bmatrix}, \Sigma\right) = \frac{1}{(2\pi)^{3/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

where $X = \begin{bmatrix} x_a \\ x_b \\ x_c \end{bmatrix}$, $\mu = \begin{bmatrix} \mu_a \\ \mu_b \\ \mu_c \end{bmatrix}$, $\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} & \Sigma_{ac} \\ \Sigma_{ba} & \Sigma_{bb} & \Sigma_{bc} \\ \Sigma_{ca} & \Sigma_{cb} & \Sigma_{cc} \end{pmatrix}$

Now Gaussians have some very useful properties:

Say we want to compute the marginal distribution of the vector $\begin{bmatrix} x_a \\ x_b \end{bmatrix}$

$$p\left(\begin{bmatrix} x_a \\ x_b \end{bmatrix}\right) = \int \mathcal{N}\left(\begin{bmatrix} x_a \\ x_b \\ x_c \end{bmatrix} \middle| \vec{\mu}, \Sigma\right) dx_c$$

$$= \mathcal{N}\left(\begin{bmatrix} x_a \\ x_b \end{bmatrix} \middle| \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}\right)$$

During Learning/Inference one often has to convert between functions of Σ^{-1} and functions of Σ.

But inverting large Σ^{-1} is expensive and as we saw above, we might not care about all values of Σ anyway.

→ Inverting partitioned matrices.

$$E^{-1} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1}+D^{-1}CMBD^{-1} \end{pmatrix}$$

where $M = (A - BD^{-1}C)^{-1}$ is known as the schur complement

If M is all that we care about, we just need to invert a matrix of size(A).; not of size(E)

→ Rank one perturbations:

Say we have already computed the inverse of a matrix $A \in \mathbb{R}^{D \times D}$

Now we change $A = A + \delta A$; where δA is of rank 1

$(A+\delta A)^{-1} = ?$

Sherman Morrison

$$(A + bUV^T)^{-1} = A^{-1} - \frac{bA^{-1}UV^TA^{-1}}{1 + bV^TA^{-1}U}$$

The cost of computing the inverse of $(A+bUV^T)^{-1}$ is $O(D^2)$ instead of $O(D^3)$.

$$= E[XX^T] - \mu_X E[X^T] - E[X]\mu_X^T + \mu_X \mu_X^T$$

$$= E[XX^T] - \mu_X \mu_X^T$$

$$= \begin{bmatrix} E[x_1^2] & E[x_1 x_2] & \cdots & E[x_1 x_N] \\ \vdots & & & \\ E[x_n x_1] & \cdots & \cdots & E[x_n^2] \end{bmatrix} - \begin{bmatrix} \mu_{x_1}^2 & \mu_{x_1}\mu_{x_2} & \cdots & \mu_{x_1}\mu_{x_N} \\ \vdots & & & \\ & & & \mu_{x_N}^2 \end{bmatrix}$$

$$= \begin{bmatrix} E[x_1^2] - E[x_1]^2 & E[x_1 x_2] - E[x_1]E[x_2] & \cdots \\ \vdots & & \\ E[x_n x_1] - E[x_n]E[x_1] & & E[x_N^2] - E[x_N]^2 \end{bmatrix}$$

$$= \begin{bmatrix} \text{Var}(x_1) & \text{Cov}(x_1, x_2) & \cdots \\ \vdots & & \\ \text{Cov}(x_n, x_1) & \cdots & \text{Var}(x_n) \end{bmatrix}$$

"The values in the covariance matrix can take any positive value between $[0, \infty]$"

Sometimes instead people like to deal with correlation matrices where values are in the bound range $[-1,+1]$.

$$corr(x,y) = \frac{cov(x,y)}{\sqrt{var[x]\,var[y]}} =$$

$$R = \begin{bmatrix} corr(x_1,x_1) & corr(x_1,x_N) \\ corr(x_N,x_1) & corr(x_N,x_N) \end{bmatrix}$$

$$= \begin{bmatrix} 1 & corr(x_1,x_N) \\ corr(x_N,x_1) & 1 \end{bmatrix}$$

'KECITATION - 02/17

— SOUMYA GHOSH

LINEAR ALGEBRA TUTORIAL

→ NOTATION:

$x \in \mathbb{R}^1 \Rightarrow$ Scalar

$X \in \mathbb{R}^{N \times 1}$ $\begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$; $x^T = [x_1 \cdots x_N]$

$X \in \mathbb{R}^{M \times N}$

$\begin{bmatrix} — & a_1 & — \\ & \vdots & \\ — & a_M & — \end{bmatrix}$ or $\begin{bmatrix} | & & | \\ b_1 & \cdots & b_N \\ | & & | \end{bmatrix}$

→ Basic Operations:

a) Vector - Vector operations.

$X, Y \in \mathbb{R}^{N \times 1}$

INNER PROD: $x^T y = \langle x, y \rangle = [x_1 \cdots x_N] \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \sum_i x_i y_i \in \mathbb{R}^1$

OUTER PRODUCT:

$$XY^T = \begin{bmatrix} x_1 y_1 & \cdots & x_1 y_N \\ \vdots & \ddots & \vdots \\ x_N y_1 & \cdots & x_N y_N \end{bmatrix}$$

Rank 1 matrix.

Let's look at an application of the outer product

Remember,

$$\begin{aligned} \text{Cov}[x_1, x_2] &= E[(x_1 - \mu_1)(x_2 - \mu_2)] \\ &= E[x_1 x_2] - \mu_1 \mu_2 \\ &= E[x_1 x_2] - E[x_1] E[x_2] \end{aligned}$$

Also $\text{Cov}[x_1, x_1] = E[x_1^2] - E[x_1]^2 = \text{var}(x_1)$.

Now, for a vector $x \in R^{N \times 1}$ $[x_1 \cdots x_D]$ we have

---

$$\begin{aligned} \text{Cov}[X] &= E[(X - \mu_X)(X - \mu_X)^T] \\ &= E[XX^T - \mu_X \mu_X^T] \\ &= E[XX^T] - \mu_X \mu_X^T \\ &= E[XX^T] - E[X] E[X]^T \end{aligned}$$

$$= \begin{bmatrix} E_x[x_1^2] & \cdots & E_x[x_1 x_N] \\ \vdots & \ddots & \\ E_x[x_N x_1] & & E_x[x_N^2] \end{bmatrix} - \begin{bmatrix} E_x[x_1]^2 & E_x[x_1]E_x[x_2] & \cdots \\ & & \end{bmatrix}$$

$$= \begin{bmatrix} E_x[x_1^2] - E_x[x_1]^2 & E_x[x_1 x_2] - E_x[x_1]E_x[x_2] & \cdots \\ & & \end{bmatrix}$$

$$= \begin{bmatrix} \text{var}(x_1) & \text{cov}(x_1 x_2) & \cdots \\ & \ddots & \\ & & \text{var}(x_N) \end{bmatrix}$$

The values in the covariance matrix lie in the range $[0, \infty]$

Aside:

$$E_x[x_1] = \int_X x_1 P(x)\, dx = \int_{x_1} \int_{x_1} x_1 P(x_1) \cdot P(x_{-1}|x_1)\, dx\, dx_{-1}$$

$$= \int_{x_1} x_1 P(x_1) \left[\int_{x_{-1}} P(x_{-1}|x_1)\, dx_{-1} \right] dx_1 = \int x_1 P(x_1)\, dx_1$$

$$= E_{x_1}[x_1]$$