ML Assignment

1) Singular value decomposition (SVD) factorizes a $m \times n$ matrix X as $X = U\Sigma V^T$, where $U \in \mathbb{R}^{m\times n}$ and $U^T U = UU^T = I$, $\Sigma \in \mathbb{R}^{m\times n}$ contains non-increasing non negative values along its diagonal and zeros elsewhere, and $V \in \mathbb{R}^{n\times n}$ and $V^T V = VV^T = I$. Given the SVD of a matrix $X = U\Sigma V^T$, what is the eigendecomposition of XX^T? (you should define an appropriate square matrix Q and diagonal matrix Λ such that $XX^T = Q\Lambda Q^{-1}$).

→ Here,
Given that,

$X = U\Sigma V^T$ where $U \in \mathbb{R}^{m\times n}$ and

$U^T U = UU^T = I$, $\Sigma \in \mathbb{R}^{m\times n}$

also, $V^T V = VV^T = I$

Now, Eigendecomposition of XX^T is given by:

$$
\begin{aligned}
XX^T &= (U\Sigma V^T).(U\Sigma V^T)^T \\
&= \cancel{(U\Sigma V^T).(U^T \Sigma^T V)} \\
&= U\Sigma(V^T V)\Sigma^T . U^T \\
&= U\Sigma(I)\Sigma^T . U^T \\
&= U . I (\Sigma\Sigma^T) . U^T \\
&= U . \Sigma^2 U^T \\
&= U\Sigma^2 U^{-1} \quad [\because UU^T = I]
\end{aligned}
$$

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

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

$U\Sigma I \Sigma^T U^T$

$U\Sigma^2 U^T$

$\rightarrow U\Sigma^2 U^{-1}$

where, U is called eigen vector and Σ^2 is called eigen values.

2)

Positive (semi-) - Definite Matrices (10 pts each part).

Let A be a real, symmetric $d \times d$ matrix. We say A is the semi-definite (PSD) if, for all $x \in \mathbb{R}^d$, $x^TAx \geq 0$. We say A is +ve definite (PD) if, for all $x \neq 0$, $x^TAx > 0$. We write $A \geq 0$ when A is PSD, and $A > 0$ when A is PD.

The spectral theorem says that every real symmetric matrix A can be expressed $A = U\Lambda U^T$, where U is a $d \times d$ matrix such that $UU^T = U^TU = I$ (called an orthogonal matrix), and $\Lambda = diag(\lambda_1, \dots, \lambda_d)$. Multiplying on the right by U we see that $AU = U\Lambda$. If we let u_i denote the ith column of U, we have $Au_i = \lambda_i u_i$ for each i. This expression reveals that the λ_i are eigenvalues of A, and the corresponding columns u_i are eigenvectors associated to λ_i.

Using the spectral decomposition, show that

a) A is PSD iff $\lambda_i \geq 0$ for each i.

b) b) A is PD iff $\lambda_i > 0$ for each i.

Hint: $U\Lambda U^T = \sum_{i=1}^{d} \lambda_i u_i u_i^T$ (you do not need to prove this. But please try to understand why this holds.

→ here,

Given, A is PSD if for all $x \in \mathbb{R}^d$, $x^TAx \geq 0$.

and A is PD if for all, $x \neq 0$, $x^TAx > 0$.

So, when, A is PSD, we write $A \geq 0$

when A is PD, we write $A > 0$.

a) A is PSD iff $\lambda_i \geq 0$ for each i. proof:

→ By spectral theorem,

$A = U\Lambda U^T$ —(a) where U is a $d \times d$ matrix

such that $UU^T = U^TU = I$ and $\Lambda = diag(\lambda_1, \dots, \lambda_d)$.

Multiplying eq^n (a) on both sides by U.

$AU = U\Lambda \underline{U^TU}$

or $AU = U\Lambda$ —(b) $[\because U^TU = I]$

At first we need to show, $\lambda_i \geq 0$ if A is PSD,

PSD properties holds that: for all $x \in \mathbb{R}^d$, $x^TAx \geq 0$

Hence, Multiplying eq^n (b) by U^T on both,

$U^TAU = U^TU\Lambda$ $[\because U^TU = I]$

$= I\Lambda$

$= \Lambda$

where, $\Lambda \geq 0$.

So, we get $U^T A U \geq 0$ [every element of Λ i.e $\lambda_i \geq 0$].

$\Lambda \geq 0$

So, $\lambda_i \geq 0$ Hence proved.

Second part

To prove that A is PSD if $\lambda_i \geq 0$.

Given that by spectral theorem,

$$A = U \Lambda U^T$$

Multiply by Y^T on both sides

$$Y^T A = Y^T U \Lambda U^T$$

Again, multiply by Y to get it in in the form of : $X^T A X$

$$Y^T A Y = Y^T U \Lambda U^T Y \quad \text{— (1)}$$

Now, Let $P = U^T Y$

then, $P^T = (U^T Y)^T$

$= Y^T U$

Then, Eqn (1) is:

$$Y^T A Y = P^T \Lambda P \quad \text{— (2)}$$

Let's assume

$$P = \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_d \end{bmatrix} \text{ and, } P^T = \begin{bmatrix} P_1, P_2 \ldots P_d \end{bmatrix}$$

also, we have $\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_d \end{bmatrix}$

$$P^T \Lambda P = \begin{bmatrix} P_1, P_2 \ldots P_d \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 \ldots & \\ 0 & & \lambda_d \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_d \end{bmatrix}$$

$$= \begin{bmatrix} P_1, P_2 \ldots P_d \end{bmatrix} \begin{bmatrix} \lambda_1 P_1 \\ \lambda_2 P_2 \\ \vdots \\ \lambda_d P_d \end{bmatrix}$$

$$= \lambda_1 P_1^2 + \lambda_2 P_2^2 + \cdots \lambda_d P_d^2$$

$$= \sum_{i=1}^{d} \lambda_i P_i^2$$

hence, Eqn (2) is: $Y^T A Y = \sum_{i=1}^{d} \lambda_i P_i^2$

CENTER FOR GLOBAL PROGRAMS & SERVICES
PHONE: 302-831-2115 FAX: 302-831-

$$Y^T A Y = \sum_{i=1}^{n} \lambda_i P_i^2$$

Since, $P_i \neq 0$, $P_i^2 \geq 0$

which gives,

If $\lambda_i \geq 0$, $Y^T A Y \geq 0$

So, If $\lambda_i \geq 0$, then A is PSD

And If $\lambda_i > 0$, $Y^T A Y > 0$ & A is PD

Hence proved

3) Maximum Likelihood estimation (15 pts).

Consider a random variable x (possibly a vector) whose distribution (density function or mass function) belongs to a parametric family. The density or mass function may be written ~~f(x;~~ $f(x;\theta)$, where θ is called the parameter, and can be either a scalar or vector. For example, in the Gaussian family, θ can be a two dimensional vector consisting of the mean and variance. Suppose the parametric family is known, but the value of the parameter is unknown. It is often of interest to estimate this parameter from observations of x.

Maximum likelihood estimation is one of the most important parameter estimation techniques. Let $X_1, \dots X_n$ be i.i.d (independent & identically distributed) random variables distributed according to $f(x;\theta)$. By independence, the joint distribution of the observations is the product.

$$\prod_{i=1}^{n} f(X_i;\theta) \quad \text{— (1)}$$

Viewed as a function of θ, this quantity is called the likelihood of θ. It is often more convenient to work with the log-likelihood

$$\sum_{i=1}^{n} \log f(X_i;\theta) \quad \text{— (2)}$$

A maximum likelihood estimate (MLE) of θ is any parameter

$$\hat{\theta} \in \arg\max_{\theta} \sum_{i=1}^{n} \log f(X_i;\theta) \quad \text{— (3)}$$

where "arg max" denotes the set of all values achieving the maximum. If there is a unique maximizer, it is called the maximum likelihood estimate. Let $X_1, \dots X_n$ be iid poisson random variables with intensity parameter λ. Determine the maximum likelihood estimator of λ.

Y I → Here,

Solⁿ is to find out the maximum likelihood estimator of the parameter of poisson distribution.

poisson's distribution is given by :

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \quad \text{where, } x = 0, 1, 2, \cdots$$

Goal: To find the maximum likelihood estimator of parameter λ from the observations:

x_i where $i = 1, \ldots, n$

x_i are independent identically distributed values

Each of x_i is a realization of random variable that has poission distribution.

Now, log likelihood function is given by:

$$l(\lambda) = \ln \prod_{i=1}^{n} f(x_i, \lambda)$$

$$= \sum_{i=1}^{n} \ln \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

$$= \sum_{i=1}^{n} \ln e^{-\lambda} + \sum_{i=1}^{n} x_i \ln\lambda - \sum_{i=1}^{n} \ln x_i!$$

$$\arg\max_{\lambda} l(\lambda)$$

$$\frac{\partial}{\partial\lambda} l(\lambda) = 0$$

$$\frac{\partial}{\partial\lambda}\left\{ \sum_{i=1}^{n}(-\lambda) + \sum_{i=1}^{n} x_i \ln\lambda - \sum_{i=1}^{n} \ln x_i! \right\} = 0$$

or, $$-n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0$$

$$\therefore \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{(which is sample mean)}$$

∘, The maximum likelihood estimator of the parameter λ for a poisson distributed random variable x is given by the sample mean of n observations.

Scanned with CamScanner

4) Unconstrained optimization

In this problem you will prove some of properties of unconstrained optimization problems. For the next two parts, the following fact will be helpful. A twice continuously differentiable function admits the quadratic expansion[1].

$$f(x) = f(y) + \langle \nabla f(y), x-y \rangle + \frac{1}{2}\langle x-y, \nabla^2 f(y)(x-y) \rangle + o(\|x-y\|^2) \quad (4)$$

where $o(t)$ denotes a function satisfying $\lim_{t \to 0} \frac{o(t)}{t} = 0$, as well as the expansion

$$f(x) = f(y) + \langle \nabla f(y), x-y \rangle + \frac{1}{2}\langle x-y, \nabla^2 f(y + t(x-y))(x-y) \rangle \quad (5)$$

for some $t \in (0,1)$.

a) Show that if f is twice continuously differentiable and x^* is a local minimizer, then $\nabla^2 f(x^*) \geq 0$, i.e. the Hessian of f is the semi-definite at the local minimizer x^*.

Hint: based on the eqn (4) we can write,

$$f(x^* + ty) = f(x^*) + \langle \nabla f(x^*), ty \rangle + \frac{t^2}{2}\langle y, \nabla^2 f(x^*) y \rangle + o(t^2)$$

) Show that if f is twice continuously differentiable, then f is convex if & only if the Hessian $\nabla^2 f(x)$ is positive semi-definite for all $x \in \mathbb{R}^d$.

Hint: the proof has two steps. Step 1. Assume that f is convex. Then prove that $\nabla^2 f(x)$ is psd. In this step, you should start with the following,

$$f(x + ty) \geq f(x) + \langle \nabla f(x), ty \rangle$$

For the left hand side, we can use equation (4). Step 2. Assume that $\nabla^2 f(x)$ is psd. Then show that f is conex. To do that, start with eqn (5)

) consider the function $f(x) = \frac{1}{2} x^T A x + b^T x + c$, where A is a symmetric $d \times d$ matrix. Derive the Hessian of f. Under what conditions on A is f convex? strictly convex?

b) → Proof : that, f is twice continuously differentiable.

→ Here,

Proof: (a)

we make the use of 2nd order Taylor series expansion eqn (4):

~~$f(x^*+ty) = f(x^*) + (\nabla f(x^*), ty) + \frac{t^2}{2}(y, \nabla^2 f(x^*) y) + o(t^2)$~~

$$f(x) = f(y) + (\nabla f(y), x-y) + \frac{1}{2}(x-y, \nabla^2 f(y)(x-y)) + o(\|x-y\|^2) \quad \text{— (4)}$$

If x^* is a local minimizes then,

The ~~to~~ second order Taylor series expansion of f at a given point $x^* \in \mathbb{R}^n$ is given by:

$$f(x) = f(x^*) + \nabla f(x^*)^T (x-x^*) + \frac{1}{2}(x-x^*)^T \nabla^2 f(x^*)(x-x^*) + o(\|x-x^*\|^2) \quad \text{— (a)}$$

where, $\lim_{x \to x^*} \frac{o(\|x-x^*\|^2)}{\|x-x^*\|^2} = 0$.

Now, $\nabla f(x^*) = 0$ because If x^* is a local minimum of f, then $\nabla f(x^*) = 0$

Given, $y \in \mathbb{R}^n$ a $t > 0$

let $x = x^* + ty$, put this into eqn (a)

we get $0 \le \frac{f(x^*+ty) - f(x^*)}{t^2} = \frac{1}{2} y^T \nabla^2 f(x^*) d + \frac{o(t^2)}{t^2}$

as $\nabla f(x^*) = 0$, Take the limit as $t \to 0$, we get

$$0 \le y^T \nabla^2 f(x^*) y$$

as y was chosen arbitrarily, $\nabla^2 f(x^*)$ is positive semi definite.

which means that the hessian of f is PSD at the local minimizes x^*

Proved

4. b) → Proof:

Given that, f is twice continuously differentiable.

To prove: f is convex if and only if Hessian $\nabla^2 f(x)$ is PSD for all $x \in R^d$.

Now, suppose that f is convex and $x \in R^d$ then by

$$f(x+ty) \geq f(x) + (\nabla f(x), ty)$$

or, $f(x+ty) \geq f(x) + t \nabla f(x)^T y$ —— (a)

for all $t \in R$.

Then, Replacing the L.H.S of the inequality (a) with its second order Taylor expansion which gives:

$$f(x) + t \nabla f(x)^T y + \frac{t^2}{2} y^T \cancel{f} \nabla^2 f(x) y + o(t^2) \geq f(x) + t \nabla f(x)^T y$$

or, $\frac{1}{2} y^t \nabla^2 f(x) y + \frac{o(t^2)}{t^2} \geq 0$

$t \to 0$ yields,

$$y^T \nabla^2 f(x) y \geq 0$$

As y was arbitrary, $\nabla^2 f(x)$ is positive semi definite.

Conversely, if $x, y \in R^d$, then by Mean value Theorem there is a $\sigma \in (0,1)$ such that

$$f(x) = f(y) + \nabla f(y)^T (x-y) + \frac{1}{2}(x-y)^T f(y_\sigma)(x-y)$$

which is eq^n (5)

where, $y_\sigma = \sigma x + (1-\sigma) y$

So,

$$f(x) \geq f(y) + \nabla f(y)^T (x-y)$$

As, $\nabla^2 f(y_\sigma)$ is PSD, f is convex. ~~by th~~

according to the eq^n: for all $x, y \in R^d$,

$$f(x) \geq f(y) + \nabla f(y)^T (x-y)$$

proved

$\rightarrow$ Here,

Given that,

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

and A is symmetric $d \times d$ matrix

To derive Hessian of f.

Let us consider a matrix as.

$$\nabla f(x) = \begin{bmatrix} \cancel{\sum_{i=1}^{d} x_i a_{i1}} \\ \cancel{\sum_{i=1}^{d} x_i a_{i2}} \\ \vdots \\ \sum_{i=1}^{d} x_i a_{id} \end{bmatrix} + \begin{bmatrix} \\ \\ \\ \end{bmatrix}$$

we have a relation in matrix notation,

$$\nabla f(x) = A^T x + Ax = (A^T + A)x$$

and, $\nabla f(x) = (A^T + A)x$

and, if A is symmetric then, $A^T = A$ so, we have,

$$(A^T + A) = (A + A) = 2A \quad \text{Hence,} \quad —(y)$$

$$\nabla f(x) = 2Ax \quad —(1)$$

Then, consider the above quadratic eq^n,

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

Using the results from eq^n(1) $^{\& (y)}$ we have,

$$\nabla f(x) = \frac{1}{2}(A^T + A)x + b \qquad \left[\because \frac{d(x^T a)}{dx} = \frac{d(a^T x)}{dx} = a^T\right.$$

$$\nabla f(x) = \frac{1}{\cancel{2}} \times \cancel{2} A x + b \qquad [\because A \text{ is symmetric}]$$

or, $\nabla f(x) = Ax + b$

$\therefore \nabla^2 f(x) = H(x) = A$;

Again, To find the conditions under which f is convex, a strictly convex

The function $f(x) := \frac{1}{2}x^T A x + b^T x + c$ is a convex function if and only if A is positive semi definite.

The function $f(x) = \frac{1}{2}x^T A x + b^T x + c$ is strictly convex if A is positive definite and, only if $A > 0$;