4. Linear Algebra Application

Linear Algebra Application

Sun Woo Lim

January 16, 2021

Norm on Vector Space

A function $||\cdot||: V \to \mathbb{R}_+$ is a norm if

1. $||x|| > 0$ if $x \neq 0$ and $||x|| = 0 \leftrightarrow x = 0$
2. $||x+y|| \leq ||x|| + ||y||$ for $x, y \in V$
3. $||\alpha x|| = |\alpha|||x||$ for $\alpha \in \mathbb{R}, x \in V$.

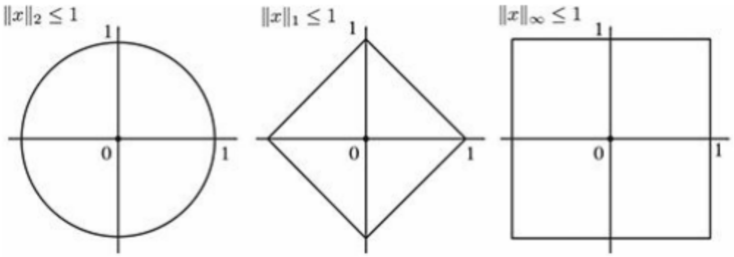
The representative example of a norm is the l_p norm : $||x||_p = (\sum |x_k|^p)^{1/p}, p = 1, 2, 3, ..., \infty$

Useful properties of l_p norms

1. $||x||_\infty := max_i |x_i| = lim_{p\to\infty} ||x||_p$
2. $\forall x \in \mathbb{R}^n$, $||x||_2/\sqrt{n} \leq ||x||_\infty \leq ||x||_2 \leq ||x||_1 \leq \sqrt{n}||x||_2 \leq n||x||_\infty$

Unit balls in the l_p norms

$B_p := \{x \in R^n : ||x||_p \leq 1\}$ is called an unit l_p norm ball.



Positive (Semi) Definiteness and Partial Order

For real, symmetric matrices (matrices we can orthogonally diagonalize as $U\Lambda U^T$),

$\mathbb{S}^n_+$ is a set of positive semidefinite real symmetric matrices.
$\leftrightarrow$ All eigenvalues $=$ diagonal entries of Λ are nonnegative

$\mathbb{S}^n_{++}$ is a set of positive definite real symmetric matrices.
$\leftrightarrow$ All eigenvalues $=$ diagonal entries of Λ are positive

Partial Order of $\mathbb{S}^n$
More generally,

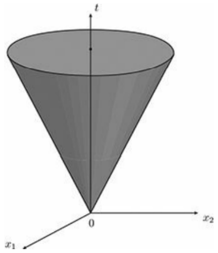
For $A, B \in S^n$, say $A \succeq B$ if $A - B \in \mathbb{S}^n_+ \leftrightarrow x^T Ax \geq x^T Bx, \forall x \in R^n$
For $A, B \in S^n$, say $A \succ B$ if $A - B \in \mathbb{S}^n_{++} \leftrightarrow x^T Ax > x^T Bx, \forall x \in R^n - \{0\}$

Why "more generally"?
$A \in \mathbb{S}^n_+$ iff $A \succeq 0$ $A \in \mathbb{S}^n_{++}$ iff $A \succ 0$

Symmetric Matrices and Cone

A "Cone" $W \subseteq \mathbb{R}^n$ is a set if for $w \in W, \alpha \geq 0$, $\alpha w \in W$

A usually dealt example of a "Cone" is a "Second-Order Cone" : $K := \{(x,t) \in \mathbb{R}^n \times \mathbb{R} : ||x||_2 \leq t\}$



PSD Matrices are cones!

S^n_{++} is the interior of the cone S^n_+

S^n is a subspace of $R^{n,n}$ while S^n_+ is not.

Matrix Norms

Matrix Norm

- $f(A) \geq 0$, and $f(A) = 0$ if and only if $A = 0$;
- $f(\alpha A) = |\alpha| f(A)$;
- $f(A+B) \leq f(A) + f(B)$.

A function $|| \cdot || : V \to \mathbb{R}_+$ is a norm if

1. $||x|| > 0$ if $x \neq 0$ and $||x|| = 0 \leftrightarrow x = 0$
2. $||x+y|| \leq ||x|| + ||y||$ for $x, y \in V$
3. $||\alpha x|| = |\alpha| ||x||$ for $\alpha \in \mathbb{R}, x \in V$.

Many of the popular matrix norms also satisfy a fourth condition called *sub-multiplicativity*: for any conformably sized matrices A, B

$$f(AB) \leq f(A) f(B).$$

Three Popular types of Matrix Norm

For $A \in \mathbb{R}^{m,n}$,

1. **Frobenius Norm** $||A||_F := (\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2)^{\frac{1}{2}} = \sum \lambda_i(A^T A) = \sum \sigma_i^2$, where σ_i's are singular values of A
2. **Nuclear Norm** $||A||_* := \sum \sigma_i$
3. **Induced Norm = Operator Norm** $||A||_p := \sup_{x \neq 0} \frac{||Ax||_p}{||x||_p}$

   $||A||_2 := \sup_{x \neq 0} \frac{||Ax||_2}{||x||_2} = \sigma_1$: largest singular value of A

   $||A||_1 := \sup_{x \neq 0} \frac{||Ax||_1}{||x||_1}$:largest absolute column sum

$||A||_2$ and Eigenvalue of A^TA

$||A||_2 := \sup_{x\neq 0} \frac{||Ax||_2}{||x||_2} = \sup_{||x||_2=1} x^T A^T Ax = \sqrt{\lambda_{max} A^T A}$

I want to prove more general result:

$\lambda_{min}(A^TA) \leq \frac{||Ax||_2^2}{||x||_2^2} \leq \lambda_{max}(A^TA)$

Basic knowledge) $A^TA \in \mathbb{S}^n_+$.

Since $A^TA \in \mathbb{S}^n$, apply spectral thm on $A^TA = U\Lambda U^T$.
$\rightarrow x^TA^TAx = x^TU\Lambda U^Tx = \tilde{x}\Lambda\tilde{x}$ where $\tilde{x} := U^Tx$

I know that $\lambda_{min}\sum\tilde{x_i}^2 \leq \sum\lambda_i\tilde{x_i}^2 \leq \lambda_{max}\tilde{x_i}^2$.
Why?) Use $A^TA \in \mathbb{S}^n_+$: All eigenvalues are nonnegative!

Using U : orthogonal : preserves length,
$\tilde{x_i}^2 = \tilde{x}^T\tilde{x} = x^Tx = ||x||_2^2$, Q.E.D.

$\lambda_{min}(A^TA) \leq \frac{||Ax||_2^2}{||x||_2^2} \leq \lambda_{max}(A^TA)$ is called **Rayleigh Quotient**.

Variational Characterization of Eigenvalues of PSD matrices

For $B \in \mathbb{S}^n_+$, its real eigenvalues satisfy :

$\lambda_1 \geq \lambda_2 ... \geq \lambda_n \geq 0.$

$\lambda_1 = \sup_{||x||_2=1} x^T Bx$

$\lambda_k = \sup_{dim(\mathbb{V})=k} \inf_{x \in \mathbb{V}, ||x||_2=1} x^T Bx = \inf_{dim(\mathbb{V})=n-k+1} \sup_{x \in \mathbb{V}, ||x||_2=1} x^T Bx.$

$\lambda_n = \inf_{||x||_2=1} x^T Bx$

$||A||_2$ and Singular value of A

$||A||_2 := \sup_{x \neq 0} \frac{||Ax||_2}{||x||_2} = \sup_{||x||_2=1} \sqrt{x^T A^T A x} = \sqrt{\lambda_{max} A^T A} = \sigma_1$
σ_1 is the biggest singular value.

What are singular values?
✓ The roots of nonzero eigenvalues of $A^T A$ are singular values of A.
✓ Since $A^T A$ is PSD, they are the roots of positive eigenvalues of $A^T A$.

✓ Important fact that $N(A^T A) = N(A)$ by definition of null space.

✓ Then, by the FTLA, $r := rank(A) = rank(A^T) = rank(A^T A) = rank(AA^T)$

✓ Thus, can align r singular values of A :

$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0.$

✓ Find a relationship with spectral decomposition of $A^T A$
: $A^T A = U \Lambda U^T$ and nonzero diagonal elements of Λ : $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r$

Sensitivity Analysis Using Condition Number $\kappa(A)$ for invertible A

Let $A \in R^{n,n}$ be an invertible matrix (Watch out! no need to be symmetric).

$\kappa(A) := \frac{\sigma_1}{\sigma_n} = ||A||_2||A^{-1}||_2$ derived from $\sigma_{max}(A^{-1}) = \frac{1}{\sigma_{min}(A)}$

Sensitivity Analysis

Let x be a solution of a linear equation $Ax = y$ for A invertible and $y \neq 0$.

Curious about what happens to x if I change y slightly by δy, in other words, $A(x + \delta x) = y + \delta y$

Solve for δx results in : $\delta x = A^{-1}\delta y$

Apply the l_2 norm and apply submultiplicativity (property of usually dealt norms) :
$||\delta x||_2 \leq ||A^{-1}||_2||\delta y||_2$

Since this only considers the absolute change of the perturbation, to consider the relative change of perturbation, use :
$||y||_2 \leq ||A||_2||x||_2 \rightarrow \frac{1}{||x||_2} \leq \frac{||A||_2}{||y||_2}$ which finally leads to:

$$\frac{||\delta x||_2}{||x||_2} \leq ||A^{-1}||_2||A||_2\frac{||\delta y||_2}{||y||_2}$$

✓ When the Matrix Condition Number is big, the linear equation $Ax = y$ has high sensitivity to perturbation in x.
✓ When $\kappa(A) \rightarrow \infty$, A is near singular, a.k.a, ill-conditioned.
✓ Orthogonal Matrices have condition number of 1 : Favored object in Numerical Analysis

Compact Form SVD

First, spectral decomposition $A^TA = V\Lambda V^T$.

$A = U_r \Sigma V_r^T$ where

1. $\Sigma = diag(\sigma_1, \sigma_2, ..., \sigma_r) \in \mathbb{S}^r_{++}$
2. $V_r \in \mathbb{R}^{n,r}$ is the alignment of the first r columns of V. $v_1, v_2, ..., v_r$ are orthonormal.
3. $U_r \in \mathbb{R}^{m,r}$ is alignment of its orthonormal columns $u_1, u_2, ..., u_r \in \mathbb{R}^m$ where $u_i := \frac{Av_i}{\sigma_i}, i = 1, 2, ..., r$.

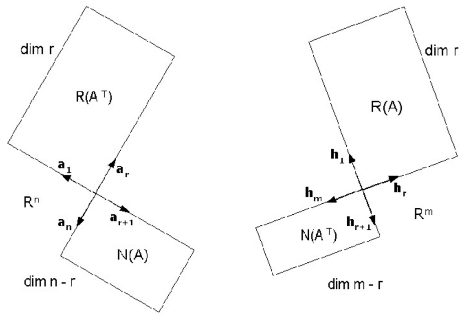
✓ No guarantee that V_r^T and U_r are orthogonal matrices! Watch out the rank r!
✓ $V_r^T V_r = I_r$ but $V_r V_r^T$ is not necessarily I_m.

Full Form SVD

$A = U\tilde{\Sigma}V^T$ where

1. $\tilde{\Sigma} = \begin{bmatrix} \Sigma & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{bmatrix}$
2. $V \in \mathbb{R}^{n,n}$ is the same V as in $A^TA = V\Lambda V^T$.
3. $U \in \mathbb{R}^{m,m}$ form an orthonormal basis of $\mathbb{R}^m$. Obtain U from $AV = U\Sigma$

✓ In Full form SVD, yes, U and V^T are orthogonal matrices.

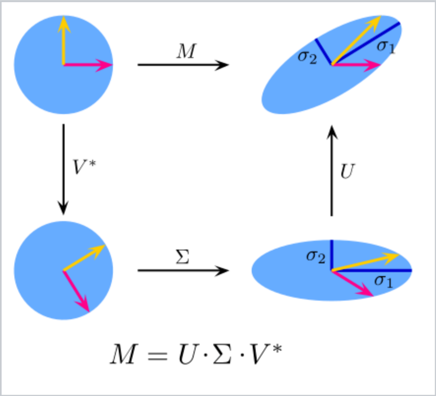


1. $\{u_1, u_2, ..., u_r\}$ form an orthonormal basis of $R(A)$
2. $\{u_{r+1}, u_{r+2}, ..., u_m\}$ form an orthonormal basis for $N(A^T)$.
3. $\{v_1, v_2, ..., v_r\}$ form an orthonormal basis for $R(A^T)$
4. $\{v_{r+1}, u_{r+2}, ..., v_n\}$ form an orthonormal basis for $N(A)$

Geometric Understanding of the Full Form SVD

$A = U\tilde{\Sigma}V^T$, $A \in \mathbb{R}^{m,n}$.
For $x \in R^n$,

1. $V^T x$ represents x in the orthonormal basis of V's column vectors.
2. $\Sigma V^T x$ represents a scaling of i^{th} component of $V^T x$ by a scaling factor σ_i.
3. $Ax = U\Sigma V^T x$ represents a linear combination of first r columns of U, having coefficients as components of $\Sigma V^T x$.



✓ Perform a compact form full form SVD using a matrix $A = \begin{bmatrix} \sqrt{\frac{3}{2}} & \sqrt{\frac{1}{2}} \\ -\sqrt{\frac{3}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix}$

First SVD application : Principal Component Analysis

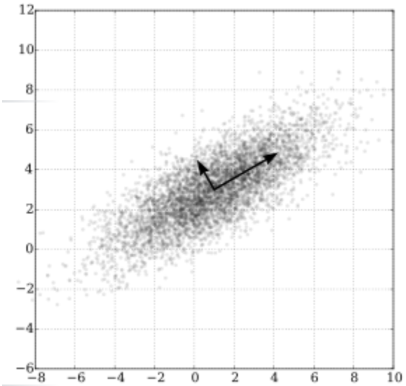
Let $X \in R^{m,n}$ be the data you want to reduce dimension from.
Let $\tilde{X} := X - \bar{X}$ be the mean centered data.

$\tilde{X} = U_r \Sigma V_r^T = U\tilde{\Sigma}V^T$: SVD on $\tilde{X}$
$\tilde{X}\tilde{X}^T = V\Lambda V^T$: Spectral Decomposition on $\tilde{X}\tilde{X}^T$

Solve $max_{z \in R^m} z^T(\tilde{X}\tilde{X}^T)z$ s.t. $||z||_2 = 1$

The optimal z is the eigenvector of $\tilde{X}\tilde{X}^T$ corresponding to the largest eigenvalue $=$ column of V corresponding to the largest singular value of $\tilde{X}$

✓ i^{th} column of $U\tilde{\Sigma} = \widetilde{X}V$ is called i^{th} principal component of X.



Second SVD application : Pseudoinverse and Least Squares

Moore - Penrose pseudoinverse of $A \in R^{m,n}$

✓ Reminder of the Compact Form / Full Form SVD

For $A \in \mathbb{R}^{m,n}$, $A = U_r \Sigma V_r^T = U \tilde{\Sigma} V^T$.

$A^\dagger := V_r \Sigma^{-1} U_r^T = V \tilde{\Sigma}^\dagger U^T$, $A^\dagger \in \mathbb{R}^{n,m}$

$\tilde{\Sigma}^\dagger := \begin{bmatrix} \Sigma^{-1} & 0_{r,m-r} \\ 0_{n-r,r} & 0_{n-r,m-r} \end{bmatrix}$

Properties of Pseudoinverse of A

1. $AA^\dagger = U_r \Sigma V_r^T V_r \Sigma^{-1} U_r^T = U_r U_r^T$
2. $A^\dagger A = V_r \Sigma^{-1} U_r^T U_r \Sigma V_r^T = V_r V_r^T$
3. $AA^\dagger A = U_r U_r^T U_r \Sigma V_r^T = U_r \Sigma V_r^T = A$
4. $A^\dagger A A^\dagger = V_r V_r^T V_r \Sigma^{-1} U_r^T = V_r \Sigma^{-1} U_r^T = A^\dagger$

Formula of Pseudoinverse

1. If A is full column rank : $r = n \rightarrow A^\dagger A = I_n$ $\;A^\dagger = (A^T A)^{-1} A^T$: Left Inverse
2. If A is fullrow rank : $r = m \rightarrow AA^\dagger = I_m$ $\;A^\dagger = A^T (AA^T)^{-1}$: Right Inverse

Projection Matrix in Regression

Projection Matrix

I want to project a vector $y \in R^m$ onto range space of $X \in R^{m,n}$ for $r := rank(X) = n$, full column rank.

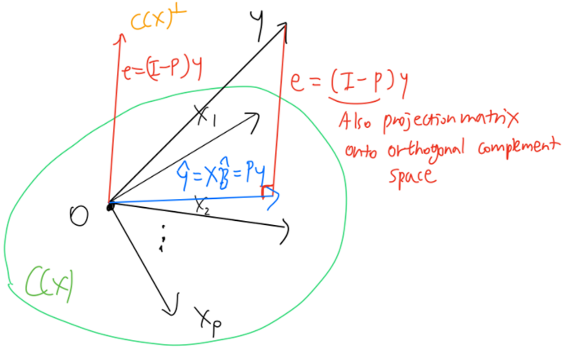
Let x_i be columns of X. Then, $x_i^T(y - X\hat{\beta}) = 0, \forall i \in \{1, ..., n\}$

In matrix form, $X^T(y - X\hat{\beta}) = 0 \leftrightarrow X^TX\hat{\beta} = X^Ty$

$\hat{\beta} = (X^TX)^{-1}X^Ty = X^\dagger y$ and $\hat{y} = X\hat{\beta} = X(X^TX)^{-1}X^Ty = Py$ for $P := X(X^TX)^{-1}X^T$

Properties of Projection Matrix

1. $P^T = P$
2. $P^2 = P$



Least Squares : Representative Example of Projection

Let the model matrix be $X \in R^{m,n}$ and the response $y \in R^m$.
Define $x_i^T \in R^n$ be the i^{th} row of X and $y_i \in R$ be the i^{th} element of y.

$min_{\beta \in R^n} ||X\beta - y||_2^2 = min_{\beta \in R^n} \sum_{i=1}^{m} (x_i^T \beta - y_i)^2$

Solving Least Square Problems using QR decomposition

In case $m \geq n$ and $rank(X) = n$,

can write $X = QR$ where
1) $Q \in R^{m,n}$ having orthonormal columns
2) $R \in R^{n,n}$ being upper triangular and invertible.

$X^T X = R^T Q^T Q R = R^T R$

Solving $X^T X \beta = X^T y \leftrightarrow R^T R \beta = R^T Q^T \leftrightarrow Rx = Q^T y$

Solve this by Back Substitution!

Types of Least Squares

Three Types of Least Squares Problem

1. Squares System : # of equations = # of unknown variables.
   $X\beta = y$. If X is full rank, $\beta^* = X^{-1}y$.
2. Overdetermined System (Used to this) : number of equations bigger than number of unknowns.
   If X is full column rank, $\beta^* = X^{\dagger}y = (X^TX)^{-1}X^Ty$
3. Underdetermined System : number of equations smaller than number of unknowns.
   Even though X is full row rank : $rank(X) = m$, $dim(N(A)) = n - m > 0$, infinite number of solutions.
   $\rightarrow$ Change this into another problem (ex : minimum norm solution)

   $min_\beta ||\beta||_2^2$ s.t. $X\beta = y$

Special Types of Least Squares

Least Squares with Equality Constraints
Find solutions to $min_{C\beta=d}||X\beta - y||_2^2$ for $X \in R^{m,n}$, $C \in \mathbb{R}^{p,n}$ and $d \in \mathbb{R}^p$

Here, assume that the problem is feasible : $\exists\tilde{\beta} \in \mathbb{R}^n$ satisfying $C\tilde{\beta} = d$
The feasible set : $\{\tilde{\beta} + Nz | z \in R^l\}$ where columns of $N \in R^{n,l}$ form a basis for $N(C)$

Vector Derivative Practice)

$\min_{x\in\mathbb{R}^n} \|Ax-b\|_2^2$ s.t. $Gx=h$, $A\in\mathbb{R}^{m\times n}$, $b\in\mathbb{R}^m$, $G\in\mathbb{R}^{p\times n}$, $h\in\mathbb{R}^p$, $\mathrm{rank}(A)=n$

"p equality constraints"

$$\begin{aligned}\mathcal{L}(x,r) &= \|Ax-b\|_2^2 + r^T(Gx-h)\\ &= (Ax-b)^T(Ax-b) + r^T(Gx-h)\\ &= (x^TA^T-b^T)(Ax-b)+r^T(Gx-h)\\ &= x^TA^TAx + (r^TG - 2b^TA)x - r^Th + b^Tb\\ &= x^TA^TAx + (G^Tr - 2A^Tb)^Tx - r^Th + b^Tb\end{aligned}$$

Using convexity of the Lagrangian w.r.t. x,

$$\nabla_x \mathcal{L}(x,r) = 2A^TAx + (G^Tr - 2A^Tb)x \overset{\text{let}}{=} 0$$

$$x^* = (A^TA)^{-1}\left(A^Tb + G^T\left(G(A^TA)^{-1}G^T\right)^{-1}\left(h - G(A^TA)^{-1}A^Tb\right)\right).$$

The original problem can be reformulated as $min_{z\in\mathbb{R}^l}||\tilde{X}z - \tilde{y}||_2^2$
for $\tilde{X} := XN$ and $\tilde{y} := y - X\tilde{\beta}$

Special Types of Least Squares

Ridge Regression

$min_{\beta \in \mathbb{R}^n} ||X\beta - y||_2^2$ has solutions $\beta^* = X^\dagger y$.
Suppose $||\beta^*||_2 = ||X^\dagger y||_2$: big.

$\rightarrow$ solve instead : $min_{\beta \in \mathbb{R}^n} ||X\beta - y||_2^2 + \lambda ||\beta||_2^2 \leftrightarrow min_{\beta \in \mathbb{R}^n} \ || \begin{bmatrix} X\beta - I_m y \\ \sqrt{\lambda} I_n \beta - 0_{n,m} \end{bmatrix} ||_2^2$

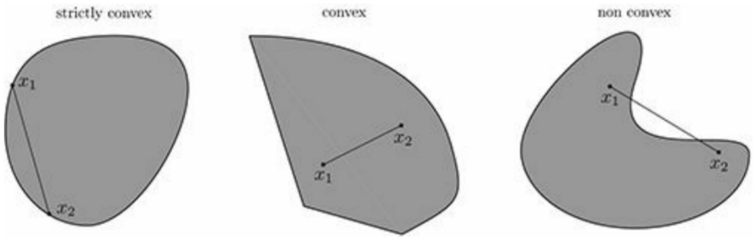
If $rank(X) = n \rightarrow$ rank($\begin{bmatrix} X \\ \sqrt{\lambda} I_n \end{bmatrix}$) = n, then ultimately get the minimizer
$\beta^* = (X^T X + \lambda I_n)^{-1} X^T y$

LASSO (Least Absolute Shrinkage and Selection Operator)

For a hyperparamter λ, Solve, $min_{\beta \in R^n} ||X\beta - y||_2^2 + \lambda ||\beta||_1$
This l_1 regularization does a variable selection by making many entries of β negligible.

Convex Set

Definition of a Convex Set



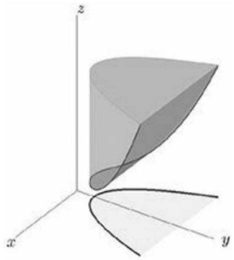
$K \subseteq \mathbb{R}^n$ is a **convex set** if $\forall x_1, x_2 \in K$, $\forall \lambda \in [0, 1]$, $\lambda x_1 + (1 - \lambda)x_2 \in K$
"The Convex Combination also is in the set".

✓ Which of the following is / are convex?

1. **Empty Set** ϕ
2. **A set of single point** $\{x_0\}$
3. $\{z \in R^n : ||z - z_0||_2 \leq \epsilon\}$ **for some** $\epsilon > 0$
4. $\{z \in R^n : ||z - z_0||_2 = \epsilon\}$ **for some** $\epsilon > 0$
5. $[-2, -1] \cup [1, 2]$

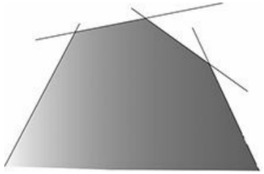
Operations preserving Convexity of a set

1. **Intersection of convex sets**
2. **Hyperplane** $\{x|a^Tx - b = 0\}$ **and Half Spaces** $\{x|a^Tx - b \leq 0\}$ **and** $\{x|a^Tx - b > 0\}$
3. **Projection of a convex set onto a hyperplane**



4. **"Convex Hull of A"** $Co(A) := \{\sum_{i=1}^{m} \lambda_i x_i | x_i \in A, \lambda_i \geq 0, \forall i, \sum \lambda_i = 1\}$
5. **"Conic Hull of A" :=**$\{\sum_{i=1}^{m} \lambda_i x_i | x_i \in A, \lambda_i \geq 0, \forall i\}$
6. **"Affine Hull of A" :=**$\{\sum_{i=1}^{m} \lambda_i x_i | x_i \in A, \sum \lambda_i = 1\}$

Q) Why is a polyhedron convex?



Convex Functions

Convex Functions defined on the domain of a convex set

For $f : R^n \to R$ defined for $x \in dom(f)$: Convex set,
f is convex if $\lambda f(x_1) + (1-\lambda) f(x_2) \geq f(\lambda x_1 + (1-\lambda) x_2)$

✓ **Properties of Convex Functions**

1. **Pointwise Supremum of convex sets is a convex function**
2. **Nonnegative linear combination of Convex Functions is a convex function**
3. $f : \mathbb{R}^n \to \mathbb{R}$ **is a convex function** $\leftrightarrow epi(f)$ **is a convex set : "Epigraph Characterization of a Convex Function"**
   $epi(f) := \{(x,t) \in \mathbb{R}^n \times \mathbb{R} : x \in dom(f), t \in \mathbb{R}, f(x) \leq t$

✓ **Iff conditions for differentiable convex functions**
For f which has an open domain and differentiable on $dom(f)$,

1. **First order (gradient) condition for convexity**
   f convex $\leftrightarrow f(y) \geq f(x) + \nabla f(x)^T (y-x), \forall x, y \in dom(f)$.
2. **Second order (Hessian) condition for convexity**
   f convex $\leftrightarrow \nabla^2 f(x) \succeq 0, \forall x \in dom(f)$. "Hessian is PSD".

Convexity of the Cost Function of Softmax Regression

Model) $P(Y=1 \mid X) = \frac{e^{w^T x}}{1+e^{w^T x}} \iff \underbrace{\ln}_{\text{log}} \left(\underbrace{\frac{P(Y=1 \mid X)}{1-P(Y=1 \mid X)}}_{\text{odds}} \right) = w^T X$ (Actually, $w^T x$ can represent bias by including "1" column)

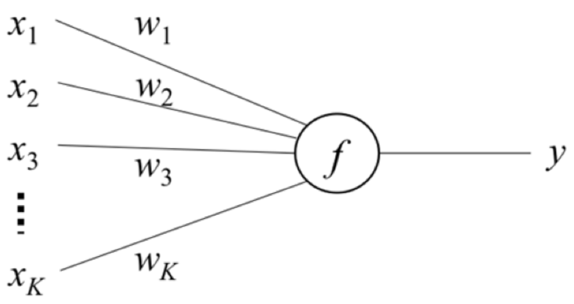
Focus on one observation of $(\vec{x}_i, y_i)$ pair.

$\text{Loss}\left(\underbrace{\widehat{P(Y_i=1 \mid X_i)}}_{\text{fitted}}, \underbrace{y_i}_{\text{true}}\right) = \begin{cases} -\log(P(Y_i=1 \mid X_i)), & \text{if } y_i=1 \\ -\log(1-P(Y_i=1 \mid X_i)), & \text{if } y_i=0 \end{cases}$

Integrate 2 cases into

$L\left(\underbrace{\widehat{P(Y_i=1 \mid X_i)}}_{\text{fitted}}, \underbrace{y_i}_{\text{true}}\right) = \underbrace{-y_i \log p(x_i; w) - (1-y) \log(1-p(x_i; w))}_{\text{entropy loss}}$

Official term : "loss": 1 case & "cost": average loss



Convexity of the Cost Function of Softmax Regression

$$\Rightarrow C(w) = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \ln p(x_i;w) + (1-y_i)\ln(1-p(x_i;w))\right]$$

use

$$p(y_i=1|x) = \frac{e^{w^Tx}}{1+e^{w^Tx}}$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \ln \exp(w^Tx)_i - y_i \ln(1+\exp(w^Tx)_i) - (1-y_i)\ln(1+\exp(w^Tx)_i)\right]$$

$$= -\frac{1}{n}\sum\left[y_i (w^Tx)_i - \ln(1+\exp(w^Tx)_i\right]$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left[y_i\left(\sum w_i x_i\right) - \ln\left(1+\exp\left(\sum w_i x_i\right)\right)\right]$$

want to find the Global minimum of this C(w) using Convexity of C(w) w.r.t. w

Convexity of the Cost Function of Softmax Regression Using the Convexity of log-sum-exp (lse)

Step 1) $C(w)$ cvx $\Longleftrightarrow -n \cdot C(w)$ concave

$$-n\,C(w) = \sum \left[y_i \underbrace{\left(\sum w_i x_i\right)}_{①} - \underbrace{\ln\left(1+\exp\left(\sum w_i x_i\right)\right)}_{②} \right]$$

Step 2) ① : linear (affine also ok) over $w \Rightarrow \boxed{\text{Concave}}$ Step 3) ② concave $\Leftrightarrow$ −② convex

pf) log-sum-exp (lse) of $\mathbb{R}^n \to \mathbb{R}$ is defined as:

$lse(\vec{w}) = \ln\left(\sum e^{w_i}\right)$, $dom(f) = \mathbb{R}^n$

Since $f := lse$ is a 2 time differentiable ftn, do a Hessian test

2. **Second order (Hessian) condition for convexity**
   f convex $\leftrightarrow \nabla^2 f(x) \succeq 0, \forall x \in dom(f)$. "Hessian is PSD".

Convexity of the Cost Function of Softmax Regression Using the Convexity of log-sum-exp (lse)

lse) "gradient" $\nabla f(w) = \frac{1}{\Sigma e^{w_i}} \begin{pmatrix} e^{w_1} \\ \vdots \\ e^{w_n} \end{pmatrix}$

2nd) "Hessian" $\nabla^2 f(w) = \frac{1}{(\Sigma e^{w_i})^2} \begin{bmatrix} e^{w_1} \\ \vdots \\ e^{w_n} \end{bmatrix} [e^{w_1} \cdots e^{w_n}]$

WTS) $\nabla^2 f(w)$ is PSD

pf) $\forall u \in \mathbb{R}^n$, $u^T \nabla^2 f(w) u = \frac{\Sigma u_i^2 e^{w_i}}{\Sigma e^{w_i}} - \frac{(\Sigma u_i e^{w_i})^2}{(\Sigma e^{w_i})^2}$

let S_i (stands for "softmax"), $S_i := \frac{e^{w_i}}{\Sigma e^{w_i}}$

$(S_1 \cdots S_n)$ Satisfies Kolmogorov Probability Axioms of Probability

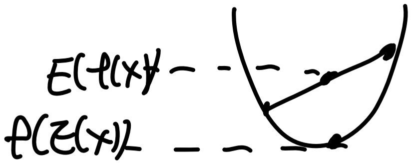
① $P(E) \geq 0$ ② $P(\Omega) = 1$ ③ For countable sequence of disjoint sets $E_1, E_2, \cdots$

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

Application : Logistic Regression and Softmax Regression

Convexity of the Cost Function of Softmax Regression Using the Convexity of log-sum-exp (lse)

$f : \text{cvx} \Rightarrow E(f(x)) \geq f(E(x))$



Then apply Jensen's Inequality :

$$\sum u_i^2 s_i \geq \left(\sum u_i s_i\right)^2 \quad " \quad E[U^2] \geq (E[U])^2 \Longleftrightarrow V[U] \geq 0$$

Thus, $u^T \nabla f(w)\, u \geq 0 \Longleftrightarrow f$: convex

Step4) ① Concave + ② concave $\Rightarrow -n\,C(w)$ concave

$\Rightarrow C(w)$ convex

Connection of the Cost Function and the Log Likelihood Function in Softmax Regression

※ Note that Log Likelihood ftn of W is

$$\therefore\ l(W) = \sum_{i=1}^{n}\left[y_i (W^T X)_i - \ln(1+\exp(W^T X)_i)\right]$$

$$= -n\,C(W)$$

"C cvx" $\Longleftrightarrow$ "l ccv"

"minimize cost" $\Longleftrightarrow$ "(log) likelihood maximization"