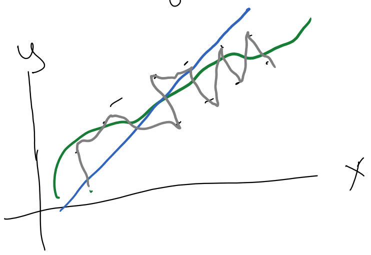
CSCC11 Tutorial

September 12, 2018 11:00 AM

T.A: Bryan Chan - email on course website



linear regression

create a function

$\hat{f}(x) = y^{*1}$ ← prediction

$\hat{g}(x) = y^{*2}$ ← prediction

how do we determine which is good? define error function

Error function to measure preformance of the model.

Errorfunction choices

- Absolute function ($F_{\overline{n}}$) → can't differentiate
- Least squares ($F_{\overline{n}}$)

Least squares:

- always non-negative ≥ 0
- Differentiable

In 1-D: we're given data $\{(x_i, y_i)\}_{i=1}^{N}$ ← input, output

To formulate error:

$$E(w) = \sum_{i=1}^{n} (\underbrace{y_i}_{\text{observation}} - \underbrace{x_i \cdot w}_{\text{prediction}})^2$$

$$= \|\vec{y} - \vec{x} w\|_2^2 \leftarrow 2^{nd} \text{ norm}$$

b is bias can be added.

find difference between observed output and input.

$= \|y - Xw\|_2$ ← 2nd norm

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Observed output and input. find distance. Motivation of cost function

In multiple dimension, we have

$(\vec{x}_i, y_i)$, $x \in \mathbb{R}^d$

$$E(\vec{w}) = \|\vec{y} - X\vec{w}\|^2_{2 \leftarrow \text{2nd norm}}, \quad X = \begin{bmatrix} \vec{x}_1^T \\ \vec{x}_2^T \\ \vec{x}_n^T \end{bmatrix}, \quad \vec{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

expanding the dot product of itself

$$= (\vec{y} - X\vec{w})^T(\vec{y} - X\vec{w})$$

$$= \underbrace{\vec{w}^T X^T X w - 2\vec{y}^T X \vec{w} + \vec{y}^T \vec{y}}_{\text{quadratic}}$$

Optimize it: Goal to minimize it. Take gradient and set it to zero, $(\vec{0})$. This will guarantee a minimum

Linear Review

1. $(AB)^T = B^T A^T$
2. $(AB)^{-1} = B^{-1} A^{-1}$ ← assuming inverse exist
3. $(A^{-1})^T = (A^T)^{-1}$
4. $|AB| = |A||B|$

   ↗ determinant(AB)

5. $|A^{-1}| = \frac{1}{|A|}$

6. A square matrix is orthogonal if every column vector

is orthogonal (check: dot product = 0) and normalized

dot product of vector by itself is 1 ($\vec{x}^T x = 1$)

Assume matrix are symmetric

Understand what orthogonal, symmetric, and invertible mean.

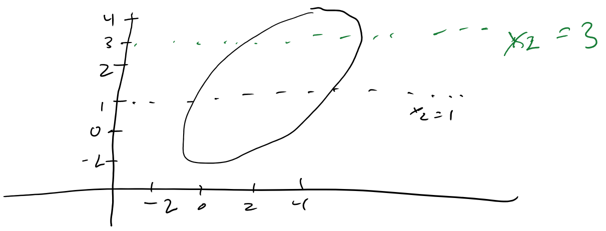
7. A square matrix "A" is not singular (invertible) if

Conditional (2D-gaussian)

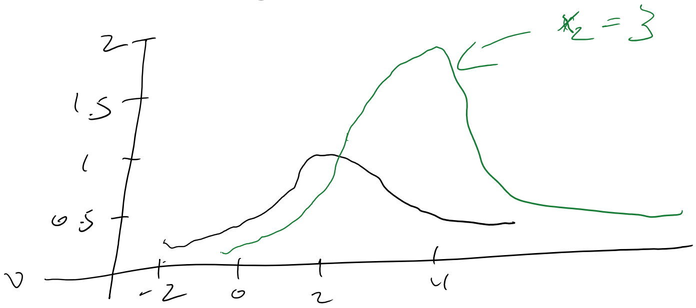
Let $\vec{x} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, $\vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$, $C = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$, $\Lambda = C^{-1} = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}$

$\therefore$ the conditional distribution of x_1 given x_2 satisfies $x_1 | x_2 \sim N(\mu_{x_1|x_2}, \Lambda_{11}^{-1})$, where where $\mu_{x_1|x_2} = \mu_1 - A_{11}^{-1} A_{12}(x_2 - \mu_2)$, note that Λ_{11}^{-1} is not simply C_{11}

For ex



$\Downarrow$



Diagonalization

Given a covariance matrix with eigen vectors $\vec{u}_1, \vec{u}_2$ and corresponding eigenvalues λ_1, λ_2

let $U = \begin{bmatrix} | & | \\ u_1 & u_2 \\ | & | \end{bmatrix}$, $S = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$,

then $CU = US$, $C\vec{u}_1 = \lambda_1 S$
$C\vec{u}_2 = \lambda_2 S$,

since U is orthonormal , orthonormal if ...

$C = USU^{-1} = USU^T$ take the inverse you got the transpose

$\Rightarrow C^{-1} = (USU^{-1})^{-1} = US^{-1}U^{-1}$

$= US^{-1}U^T$

$\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T C^{-1}(\vec{x}-\vec{\mu})\right)$, sub C^{-1} in,

$-\frac{1}{2}(\vec{x}-\vec{\mu})^T US^{-1} \underbrace{U^T(\vec{x}-\vec{\mu})}_{\text{chang of variable}})$

$y = U^T(\vec{x}-\vec{\mu}) \Rightarrow \left(-\frac{1}{2}\vec{y}^T S^{-1}\vec{y}\right)$

y is still a gaussian, with mean $= 0$, and covariance of S.

So S is diagonal matrix $\Rightarrow y_1, y_2$ are $\perp$

if covariance is diagonal then $y_1 \perp y_2$

$\Rightarrow$ Eigen values of C are the variances along the principle directions given by the eigenvectors used for the diagonalization.

Positive Definite

A $d \times d$ matrix A is positive definite iff $\forall \vec{z} \in \mathbb{R}^d - \{\vec{0}\}$, $\vec{z}^T A \vec{z} > 0$

if $\vec{z} A \vec{z} \geq 0$, then positive semi definite

why is $C > 0$?

1) Covariance is the second moment of a PDF (shifted by $\vec{\mu}$)

$$\Rightarrow C = E((\vec{x}-\vec{\mu})(\vec{x}-\vec{\mu})^T) \quad , \qquad E(a^n)$$

If C is full rank (invertible) then

$$\vec{z}^T C \vec{z} = \vec{z}^T E((\bar{x}-\bar{\mu})(\bar{x}-\bar{\mu})^T)\vec{z}$$

$$= E(z^T(x-\bar{\mu})(\bar{x}-\bar{\mu})^T z)$$

$$= E(u^2), \quad u = (\bar{x}-\bar{\mu})^T$$

$$> 0$$

mean will always be positive because of the square.

$\Rightarrow C$ is positive definite.
if $C > 0$, eigen > 0

Why?

$$C\vec{z} = \lambda \vec{z} \quad , \quad \bar{z} \text{ eigen vector (unit norm)}$$

$$\Leftrightarrow z^T C \bar{z} = \lambda > 0 \quad \bar{\lambda} \text{ eigen value}$$

$$\therefore \lambda > 0$$

Block Matricies (Reference: matrix identity)

A block matrix is a matrix that is interpreted as having been breaking into sections called block(s).

$$U = \left[\begin{array}{c|c} U_1 & U_2 \end{array} \right]$$

$$\qquad 2\times 1 \qquad 2\times 1$$

Useful in higher dimensions, use it do marginalize and condition

$p(x_1, x_2)$

$$p(x_1) = \int p(x_1, x_2)\, dx_2$$
$$p(x_2) = \int p(x_2, x_1)\, dx_1$$

For the case of gaussians
Covariance

$$\Sigma = \begin{bmatrix} A & B \\ C & D \end{bmatrix}, \quad A \in \mathbb{R}^{m \times m}, \quad D \in \mathbb{R}^{n \times n}$$

$$B = C^T$$

if $\Sigma = \begin{bmatrix} A & 0 \\ 0 & D \end{bmatrix}$ then $\Sigma^{-1} = \begin{bmatrix} A^{-1} & 0 \\ 0 & D^{-1} \end{bmatrix}$

$$\det = \det(A)\det(D)$$

Regression - Quiz

helpful tips:

1 generate noiseless data set

create own data set/training set

penalty for efficiency.

↳ around 5 seconds

Baye's Rule

$$\underbrace{P(\vec{w} \mid D)}_{\text{posterior}} = \underbrace{P(D \mid \vec{w})}_{\text{likelihood}} \times \frac{\overbrace{P(\vec{w})}^{\text{prior}}}{\underbrace{P(D)}_{\text{evidence}}}$$

$$, \quad \int P(D, w)\,dw = \int P(D|w)P(w)\,dw$$

$\vec{w}$ parameter of given model

D = training data.

$$\to \quad P(\vec{w} \mid D, M) = \frac{P(D \mid \vec{w}, M) \times P(\vec{w} \mid M)}{P(D \mid M)}$$

We care about different models.

In a1, we are trial and erroring. But there is a more rigorous way using Baye's Rule.

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)} \leftarrow \text{Constant}$$

→ Maximize posterior (give the "best" model)

Ex/ Estimating Gaussian Distribution

Suppose we are learning a gaussian distribution from n training data $\{\vec{x}_i\}_{i=1}^N$ and we want to know the best parameter $(\vec{\mu}, \Sigma)$ for this distribution.

We already have the likelihood $= P(\vec{X}_{i=N} \mid \vec{\mu}, \Sigma)$

$$= P(\underbrace{\vec{x}_1, \vec{x}_2, \ldots \vec{x}_N}_{\text{iid}} \mid \vec{\mu}, \Sigma)$$

$$= \prod_{i=1}^{N} P(\vec{X}_{i=N} \mid \vec{\mu}, \Sigma)$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} \exp\left(-\frac{1}{2} (\vec{x}_i - \vec{\mu})^T \Sigma^{-1} (\vec{x}_i - \vec{\mu})\right)$$

$$, \quad \text{note } (2\pi)^{D} \leftarrow D \text{ is Dimension}$$

⇒ Because we restricted M = Gaussian, we only need to maximize the likelihood (MLE).

$\Rightarrow$ Maximizing the above is complicated so instead, take the negative log-likelihood, minimizing it

$$\to L(\bar{\mu}, \Sigma) = -\ln\left(p(\bar{x}_{1:N} \mid \bar{\mu}, \Sigma)\right)$$

$$= -\sum_{i=1}^{N} p(\bar{x}_i \mid \bar{\mu}, \Sigma)$$

$$= \sum_{i=1}^{N}\left[\frac{(\bar{x}_i - \bar{\mu})^T \Sigma^{-1} (\bar{x}_i - \bar{\mu})}{2}\right] + \frac{N}{2}\ln|\Sigma| + \frac{ND}{2}\ln(2\pi)$$

$$\mu^* = \underset{\mu}{\operatorname{argmax}}\; p(\bar{x}_{1:N} \mid \bar{\mu}, \Sigma) = \underset{\mu}{\operatorname{argmin}}\; L(\bar{\mu}, \Sigma)$$

$$\Sigma^* = \underset{\Sigma}{\operatorname{argmax}}\; p(\bar{x}_{1:N} \mid \bar{\mu}, \Sigma) = \underset{\Sigma}{\operatorname{argmin}}\; L(\bar{\mu}, \Sigma)$$

can solve them by

$$\frac{\partial L}{\partial \bar{\mu}} = 0 \qquad \frac{\partial L}{\partial \Sigma} = 0$$

$$\bar{\mu}^* = \frac{1}{N} \sum_{i=1}^{N} \bar{x}_i$$

$$\Sigma^* = \frac{1}{N}(\bar{x}_i - \bar{\mu}^*)(\bar{x}_i - \bar{\mu}^*)^T$$

Entropy & Information Theory

- Entropy measures uncertainty of a distribution.
- Entropy is a key measure of uncertainty associated with a r.v.
- Use entropy for decision tree, $\underbrace{\text{K-L divergence}}_{\text{common measure of distance or ML.}}$, cross entropy

- Give a discrete r.v y, which takes on K values, entropy is defined as $h = -\underbrace{E_{p(c)}}_{\text{expectation}}[\log p(c)] = -\sum_{c=1}^{K} p(c)\log(p(c))$

Say we have $K=8$, $p(c=i) = \frac{1}{8}\; \forall\, i = 1, \dots, 8$

$$H = -8\left(\frac{1}{8}\log\left(\frac{1}{8}\right)\right) = 3$$

minimum bound

This is used for data compression too.