Handwritten Mathematical Analysis

① Linear Regression:

Dataset - 435742 rows, 13 columns

X (input matrix) - (435742, 4) ← pollutant concentrations of SO_2, NO_2, RSPM & SPM

y (output label vector) - (435742, 1) ← AQI

Linear Regression equation -

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$$

where $X = [x_1, x_2, \dots, x_N]$, $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{bmatrix}$

$5 \times N$ ← $1+4$ ← dummy variable

$$y = X^T \theta \quad \text{- (vectorized notation)}$$

(435742×5) (5×1)

$$P(y \mid X, \theta) = \prod_{n=1}^{N} P(y_n \mid x_n, \theta)$$

(N = total no of samples)

$$= \prod_{n=1}^{N} \mathcal{N}(y_n \mid x_n^T \theta, \sigma^2)$$

$x_n \in \mathbb{R}^4$

Taking negative log-likelihood,

$$-\log P(y \mid X, \theta) = -\log \prod_{n=1}^{N} \mathcal{N}(y_n \mid x_n^T \theta, \sigma^2)$$

$$= -\sum_{n=1}^{N} \log \mathcal{N}(y_n \mid x_n^T \theta, \sigma^2)$$

in LR model; $y = x^T \theta + \epsilon$, $\quad \epsilon \sim \mathcal{N}(0, \sigma^2)$

Hence, $\log p(y_n | x_n, \theta) = \frac{-1}{2\sigma^2}(y_n - x_n^T\theta)^2 + \text{constant}$

Hence, the negative log likelihood turns out to be, :

$$L(\theta) = -\log(p(Y|X,\theta)) = \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - x_n^T\theta)^2$$

$$L(\theta) = \frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta) = \frac{1}{2\sigma^2}\|y - X\theta\|^2$$

$X \rightarrow [x_1, \ldots, x_N]^T \in \mathbb{R}^{N \times D}$

where $N = 435742$, $D = 4$ (features)

$y \rightarrow [y_1, \ldots, y_N]^T \in \mathbb{R}^{N \times 1}$

Taking derivative of log likelihood & equating it to 0 to obtain θ_{ML}. (optimal parameter) :$\Rightarrow$

$$\frac{dL}{d\theta} = \frac{d}{d\theta}\left(\frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta)\right)$$

$$= \frac{1}{2\sigma^2}\frac{d}{d\theta}(y^Ty - 2y^TX\theta + \theta^TX^TX\theta)$$

$$= \frac{1}{\sigma^2}(-y^TX + \theta^TX^TX) \in \mathbb{R}^{1 \times D}$$

$$\frac{dL}{d\theta} = 0^T \Leftrightarrow \theta_{ML}^T X^TX = y^TX$$

$$\Leftrightarrow \theta_{ML}^T = y^TX(X^TX)^{-1}$$

MLE Equation $\rightarrow$ $\boxed{\theta_{ML} = (X^TX)^{-1}(X^Ty)}$

for Non linear transformation (Polynomial Regression)

$$y = \phi(x)^T \theta$$

where $\phi(x) = [x^0 \ x^1 \ x^2 \dots x^K]^T$

on applying the Φ matrix,

$$\Phi = \begin{bmatrix} \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_{K-1}(x_1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_N) & \cdots & \cdots & \phi_{K-1}(x_N) \end{bmatrix} \in \mathbb{R}^{N \times K}$$

K = degree of polynomial

x_n = D dimensional vector

$\phi(x_n) : \mathbb{R}^D \to \mathbb{R}$ and $\Phi_{in} = \phi_i(x_n)$

$\cancel{\Phi(x_n) : \mathbb{R}^D \to \mathbb{R}^K}$ $\quad i \in (0, K)$

In our case, after applying polynomial Regression, we get,

$$y = \theta_0 + \theta_1(x_1 + x_2 + x_3 + x_4) + \theta_2(x_1^2 + x_2^2 + x_3^2 + x_4^2) + \dots$$

for each point (x, y)

Hence, MLE equation turns out to be

$$\boxed{\theta_{ML} = (\Phi^T \Phi)^{-1} (\Phi^T y)}$$

$\Phi \in \mathbb{R}^{N \times K}$, $y \in \mathbb{R}^{N \times 1}$

degree of polynomial

total no of points

Maximum A Posteriori Estimation

To overcome the problem of overfitting in MLE estimation, we need prior distribution about what parameter values are plausible.

According to Bayes Theorem,

$$p(\theta \mid X, y) = \frac{p(Y \mid X, \theta)\, p(\theta)}{p(Y \mid X)}$$

to find the MAP estimate,

$$\log(p(\theta \mid x, Y)) = \log p(Y \mid x, \theta) + \log p(\theta) + \text{constant}$$

The gradient of the -ve log posterior with respect to θ is

$$-\frac{d(\log p(\theta \mid x, Y))}{d\theta} = -\frac{d}{d\theta}(\log p(Y \mid x, \theta)) - \frac{d}{d\theta} \log p(\theta)$$

after solving, we get,

$$-\log p(\theta \mid X, y) = \frac{1}{2\sigma^2}(y - \Phi\theta)^T(y - \Phi\theta) + \frac{1}{2b^2}\theta^T\theta + \text{constant}$$

considering Guassian prior $p(\theta) = N(0, b^2 I)$ on the parameters θ.

$$-\frac{d}{d\theta}(\log p(\theta \mid X, y)) = \frac{1}{\sigma^2}(\theta^T \Phi^T \Phi - y^T \Phi) + \frac{1}{b^2}\theta^T$$

- Support vector Machines.

On simplification, we get

$$\frac{1}{\sigma^2}\left(\theta^T\Phi^T\Phi - y^T\Phi\right) + \frac{1}{b^2}\theta^T = 0^T$$

$$\theta^T\left(\frac{1}{\sigma^2}\Phi^T\Phi + \frac{1}{b^2}I\right) - \frac{1}{b^2}y^T\Phi = 0^T$$

$$\theta^T = y^T\Phi\left(\Phi^T\Phi + \frac{\sigma^2}{b^2}I\right)^{-1}$$

$$\# \quad \boxed{\theta_{MAP} = \underbrace{\left(\Phi^T\Phi + \frac{\sigma^2}{b^2}I\right)}^{-1}\Phi^T y} \quad \#$$

MAP estimation optimal parameter

$\Phi \in \mathbb{R}^{N\times K}$, $y \in \mathbb{R}^{N\times 1}$

$N \rightarrow$ total no of datapoints

$K \rightarrow$ degree of polynomial.

The extra term ensures that it is symmetric & strictly positive definite ~~kernel~~ (also works as an effect of Regularizer)

Principal Component Analysis

$X^{N \times D}$ = original dataset (feature matrix)

$N = 434752$ datapoints, $D = 4$ features

Aim → to calculate $\tilde{X}^{N \times D}$ that has similar dimensions as that of original dataset (feature matrix) but which have a significantly lower intrinsic dimensionality.

↳ First we normalize the dataset,

$$X_{norm} = \frac{X - \mu}{\sigma}, \quad \text{where } \mu = \frac{1}{N} \sum_{i=1}^{n} X_i$$

"mean"

σ = standard dev of points in X

dataset →

$X_{norm}^T = \{x_1, \ldots, x_N\}$;

$x_n \in \mathbb{R}^D$ & with mean 0.

→ covariance matrix →

$$S = \frac{1}{N} \sum_{n=1}^{N} x_n x_n^T = \frac{1}{N} X_{norm}^T X_{norm}$$

dimension of covariance matrix → (4, 4)

Now for principal components, we need to find eigenvalues & eigenvectors of the covariance matrix S.

$$S\vec{v} = \lambda \vec{v} \quad \text{—— (1)}$$

where $\vec{v}$ → eigenvector &

λ → corresponding eigenvalue of S.

$$(S\vec{v} - \lambda\vec{v}) = 0$$

$$\vec{v}(S - \lambda I) = 0 \quad \text{—— (2)}$$

$(S-\lambda I)$ has to be non invertible,

Hence, $\det(S-\lambda I)=0$ — ③

since, S is a 4×4 matrix,

on solving eqn ③, we get 4 different values of λ.

Using $S\vec{v}=\lambda\vec{v}$,

for each λ, we get a column vector $\vec{v}$.

~~Higher~~

Hence, we get 4 eigenvalues of S and 4 correspondig eigen vectors of S.

Higher the value of λ, higher will be the variance contributed by that particular eigenvector,

Hence, we sort the eigenvectors corresponding to decreasing order of eigenvalues

$$\text{eigenvec} = [\vec{v_1}\ \ \vec{v_2}\ \ \vec{v_3}\ \ \vec{v_4}]$$

where $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4$.

(decreasing order)

$B =$ matrix that takes top M eigen vectors

where $M \rightarrow$ lower dimensional space $= 2$ (in our case)

$$B=[\vec{v_1}\ \ \vec{v_2}]$$

we then calculate projection matrix P,

$$P = BB^T$$

where $\dim(B) = 4\times M$

$\dim(P) = 4\times4$.

$$X_{reconstruct} = \underbrace{BB^T}_{\substack{\text{projection} \\ \text{matrix } P \\ (4\times4)}} \cdot \underbrace{X_{norm}^T}_{\substack{(4 \times N) \\ \text{(total data points)}}}$$

$$\underbrace{X_{reconstruct}}_{(N \times 4)} = \underbrace{X_{reconstruct}^T}_{(4 \times N)}$$

This matrix is reconstructed feature matrix whose dimension in our case = (435742, 4) (same as X_{norm}) but with lower intrinsic dimensionality.

Further, we calculate y (AQI values) from this Reconstructed feature matrix & compare the new AQI values with the actual AQI values.

- Support vector Machines.

(handwritten Mathematical Analysis)

→ The equation of hyperplane in 'M' dimension is —

$$y = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + \ldots$$

$$= w_0 + \sum_{i=1}^{m} w_i x_i = w_0 + w^T X$$

$$= \underset{\substack{\uparrow \\ \text{biased term}}}{b} + w^T X$$

→ Hard Margin SVM :

we conclude, for any point x_i,

if $y_i(w^T x_i + b) \geq 1$ then $x_i \rightarrow$ correctly classified

else $x_i \rightarrow$ incorrectly classified.

If point are linearly seperable then only our hyperplane is able to distinguish between them and if any outlier is introduced then it is not able to seperate them. so these type of SVM is called hard margin SVM.

→ Soft margin SVM :

If the points are not linearly seperable then new slack variable is introduced (ξ_i) ~~which is called~~ ~~ξ~~.

new equation:

$$y_i(w^T x_i + b) \geq 1 - \xi_i$$

if $\xi_i = 0$, points ~~can~~ be correctly classified else if $\xi_i > 0$, points are incorrectly classified.

means ξ_i = error term & the average error is,

$$\text{avg error} = \frac{1}{n}\sum_{i=1}^{n} \xi_i$$

Hence, our mathematical objective is,

$$\underset{w,b}{\text{minimize}} \ \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \xi_i$$

such that $y_i(w^T x_i + b) \geq 1 - \xi_i$, for all $i = 1, 2, \ldots, n$

→ <u>Dual form SVM:</u>

with SVM, we can seperate each data point by projecting it onto a higher dimension.

alternate method is dual form SVM which uses langragian's multiplier to solve the constraints optimization problem.

$$\underset{\alpha}{\text{maximise}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i^T x_j)$$

subject to $\alpha_i \geq 0$ for all $i = 1, 2, \ldots, n$ &

$$\sum_{i=1}^{n} \alpha_i y_i = 0.$$

If $\alpha_i > 0$ then $x_i \rightarrow$ support vector &

when $\alpha_i = 0$ then $x_i \rightarrow$ not a support vector.

→ Kernel trick

new equation with Kernel function

$$\underset{\alpha}{\text{maximise}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \underline{K(x_i^T x_j)}$$

Kernel funcn

linear kernel $K(x_1, x_2) = x_1^T x_2$.

Poly kernel $K(x_1, x_2) = (a + x_1^T x_2)^b$

gaussian kernel $K(x_1, x_2) = \exp(-\gamma \|x_1 - x_2\|^2)$

Guassian mixture model

↳ clustering algorithm

$$p(x|\mu,\Sigma,\pi) = \sum_{j=1}^{n} \pi_j \mathcal{N}(x|\mu_j,\Sigma_j)$$

subject to $\sum_{j=1}^{n} \pi_j = 1$.

↑ weighted sum of different guassian distributions

pdf of guassian distribution,

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

for multivariate guassian distribution,

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^K |\Sigma_j|}} \exp\left(-\frac{(x_i-\mu_j)^T \Sigma_j^{-1} (x_i-\mu_j)}{2}\right)$$

Expectation-Maximization algorithm is applied as the parameters cannot be estimated in the closed form.

① ~~E step:~~

We know, $p(X|\theta) = \prod_{n=1}^{N} p(x_n|\theta)$

and

$$p(x_n|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)$$

$$P(X|\theta) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)$$

$$\log(P(X|\theta)) = \underbrace{\sum_{n=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)\right)}_{\text{log likelihood } \mathcal{L}}$$

To find optimal parameters μ_k, Σ_k, π_k,

$$\frac{d\mathcal{L}}{d\mu_k} = 0^T, \quad \frac{d\mathcal{L}}{d\Sigma_k} = 0, \quad \frac{d\mathcal{L}}{d\pi_k} = 0 \quad \text{—} \quad (1)$$

Responsibility → $$r_{nk} = \frac{\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}$$

↑ Responsibility of kth mixture component for the nth data point

on applying eqn (1), we get

$$\mu_k^{new} = \frac{\sum_{n=1}^{N} r_{nk} x_n}{\sum_{n=1}^{N} r_{nk}} \quad \text{—} \quad (A)$$

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (x_n - \mu_k)(x_n - \mu_k)^T$$

where $N_k = \sum_{n=1}^{N} r_{nk}$

$$\pi_k^{new} = \frac{N_k}{N}, \quad k = 1, 2, \ldots, K.$$

N → no of datapoints

Scanned with CamScanner