11/11/17 07:55

Fazer as listas de Machine learning:

Questão 1 -

Modelo de regressão linear:

$D = \{(\vec{x}_i, y_i)\}_{N}\}$ → N pontos $(\vec{x}_i, y_i)$

↳ x_i = observado, dimensão "D" ↳ MATRIX

↳ Então podemos definir uma matriz X, tal que: $\{x_{ij} \in X \mid x_{ij}$ é o valor do feat. j do pt. $i\}$

↳ onde $1 \leq i \leq N$ e $1 \leq j \leq D$

DEFINIÇÕES

→ Eu tenho um modelo com parâmetros "θ", LIKELIHOOD

quais são os parametros que eu devo escolher!

↳ Se "p(D|θ)" é a probabilidade dos dados terem sido gerados por um determinado modelo (θ), seria bom escolher os paramentros tal que:

$\hat{\theta} = \arg\max_{\theta} p(D|\theta)$, ou seja, encontrar o MLE (MAXIMUM LIKELIHOOD ESTIMATION)

→ Por hipótese: os valores de "y" são normalmente distribuídos entorno de um valor (μ_i) que depende, linearmente, de x_i. Daí:

$p(D|\theta) = \prod p(y_i | \vec{w}_i, \vec{x}_i)$ → w_i: são os pesos

↳ erros são independentes (por hipótese)

→ Da onde $p(D|\theta) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{(y_i - \mu_i)^2}{2\sigma^2}\right)}$

~~Aplicando log,~~ $\log(p(D|\theta)) = -\log(\sigma\sqrt{2\pi})$

* Inicialmente temos:

$D = \{(\vec{x}_i, y_i) : N\}$ N pontos "i"

D → ↳ dados observados (dimensão D)

↳ $\vec{x}_i \in X$ tal que: $\{x_{ij} \in X\}^D$ MATRIZ

Aonde "x_{ij}" é o valor do ponto "i" do feature "j".

* Modelo: $\bar{y}_i = w_1 x_{i1} + w_2 x_{i2} + \ldots + w_j x_{ij}$

Onde: $\bar{y}_i = \mu_i$, para, $N_{\vec{x}_i}(\sigma_i, \mu_i)$ é dist. de "y" para o ponto $\vec{x}_i$.

ou seja: ↳ ERRO NORMALMENTE DISTRIBUÍDO

Hip. 1: $y_i(x_i) = \bar{y}(x_i) + \epsilon(x_i)$

Outro aspecto importante do modelo é que os resultados ("y_i") são independentes entre si. (Hip. 2).

* Problema: Eu tenho um modelo, esse modelo tem uma série de parâmetros desconhecidos. Quais os valores dos parâmetros que devo escolher?

↳ Eu sei que para um conjunto de parâmetros é possível calcular a probabilidade dos dados serem esplicados pelo modelo (Likelihood = P(D|θ)). Parece ser uma boa idéia buscar quais os parâmetros que maximizam essa probabilidade, ou seja:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ p(D|\theta) \qquad (1)$$

Pela "Hip.1": $P(y_i|\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-\mu_i)^2}{2\sigma^2}}$

Pela "Hip.2": $p(D|\theta) = \prod^{N} P(y_i|\theta) \Rightarrow$ Como:

$\underset{\theta}{\operatorname{argmax}}\ p(D|\theta) = \underset{\theta}{\operatorname{argmax}} \log(p(D|\theta)) = \hat{\theta} \Rightarrow$

$\Rightarrow \hat{\theta} = \underset{\theta}{\operatorname{argmax}} \left[\log\left(\prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - \vec{w}\vec{x}_i)^2}{(2\sigma^2)}}\right)\right] \Rightarrow$

$\Rightarrow \hat{\theta} = \underset{\theta}{\operatorname{argmax}} \left[-\frac{N}{2}(2\sigma^2\pi) - \frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^{N}(y_i - \vec{w}\vec{x}_i)^2}_{RSS}\right]$

Essa manipulação é interessante porque reduz a minimização de um produtório a algo mais simples (minimização somatório)

Nesse caso podemos definir que:

max. log likelihood = min. neg loglikelihood (NLL)

e $\boxed{\text{min. NLL} = \text{min RSS}}$

Ou é é por isso (pelo MLE, Hip.1 e Hip.2) que faz sentido buscar o modelo que minimiza o RSS.

$\Rightarrow$ Se definirmos (por conveniência):

$NLL(w) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \vec{w}\vec{x}_i)^2 \Rightarrow$

$NLL(x) = \frac{1}{2}(\vec{y} - [x]\vec{w})^T(y - Xw)$ FORMA VETORIAL

$\Rightarrow NLL(w) = \frac{1}{2}(\vec{y} - [x]\vec{w})^T(\vec{y} - [x]\vec{w}) =$

$= \frac{1}{2}(y^Ty - y^T[x]w) - (([x]\vec{w})^T\vec{y} + ([x]\vec{w})^T([x]w))$

$= \frac{1}{2}y^Ty - \frac{1}{2}\underbrace{(\vec{w}^T[x]^T y)^T}_{\text{Dimensão 1.1}} - \frac{1}{2}\underbrace{\vec{w}^T[x]^T\vec{y}}_{\text{Dimensão 1}} + \frac{1}{2}\vec{w}^T[x]^T[x]w =$

$= \frac{1}{2}\vec{y}^T\vec{y} - \vec{w}^T[x]^T\vec{y} + \frac{1}{2}\vec{w}^T[x]^T[x]\vec{w} = NLL(w) \Rightarrow$

$\Rightarrow \frac{\partial NLL}{\partial w} = \cancel{\frac{\partial(\frac{1}{2}\vec{y}^T\vec{y})}{\partial w}}^{0} - \frac{\partial \vec{w}[x]^T\vec{y}}{\partial w} + \frac{1}{2}\frac{\partial \vec{w}^T[x]^T[x]w}{\partial w} \Rightarrow$

$\Rightarrow \frac{\partial NLL}{\partial w} = -[x]^T\vec{y} + \frac{1}{2}(([x]^T[x])^T + [x]^T[x])w \Rightarrow$

$\Rightarrow \frac{\partial NLL(w)}{\partial w} = -[x]^Ty + [x]^T[x]w \Rightarrow$

$\Rightarrow \frac{\partial NLL(w)}{\partial w} = 0 \Rightarrow [x]^T[x]w = [x^T]y \Rightarrow$

$$\Rightarrow \boxed{\hat{\theta} = \underset{\text{MODELO LINEAR}}{\hat{w}} = ([x]^T[x])^{-1}[x]^Ty}$$

2011/17
01:57

Modelo Gaussiana multivariado:

$f(\vec{x}) = \mathcal{N}(\vec{x} \mid \theta) \longrightarrow$ Normal multivariado depende $\vec{\mu}$ e $[\sigma]$

$\Rightarrow$ função de probabilidade no ponto $\vec{x}$

$\Rightarrow$ conjunto de N pontos "x_i"

$\Rightarrow \mathcal{L}(\theta \mid X) = \prod_{i=1}^{N} \mathcal{N}(\vec{x}_i \mid \theta) \Rightarrow$

$\Rightarrow \hat{\theta} = \underset{\theta}{\arg\max}\, \mathcal{L}(\theta \mid X) = \underset{\theta}{\arg\max} \log(\mathcal{L}(\theta \mid X)) \Rightarrow$

$\Rightarrow \log(\mathcal{L}(\theta \mid X)) = \sum_{i=1}^{N} \log(\mathcal{N}(\vec{x}_i \mid \vec{\mu}_y, [\sigma])) \Rightarrow$

$\Rightarrow \log(\mathcal{L}(\theta \mid X)) = \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi[\sigma]}} e^{-\frac{1}{2}(\vec{x}_i - \vec{\mu}_y)^T [\sigma]^{-1} (\vec{x}_i - \vec{\mu}_y)} \right)$

$\Rightarrow \log(\mathcal{L}(\theta \mid X)) = -\frac{N}{2}\sqrt{2\pi[\sigma^2]} - \frac{1}{2}\sum_{i=1}^{N} (\vec{x}_i - \vec{\mu}_y)^T [\sigma]^{-1} (\vec{x}_i - \vec{\mu}_y)$

$\Rightarrow$ Seja $f(\theta) = -\frac{N}{2}\sqrt{2\pi[\sigma^2]} - \frac{1}{2}\sum_{i=1}^{N} (\vec{x}_i - \vec{\mu}_y)^T [\sigma]^{-1} (x_i - \mu_y) \Rightarrow$

$f'(\theta^*) = 0 \Rightarrow \theta^* = \hat{\theta} \Rightarrow$

$\Rightarrow \frac{\partial f}{\partial \mu} = -\frac{1}{2}\sum_{i=1}^{N} \frac{\partial (x_i - \mu_y)^T [\sigma]^{-1} (x_i - \mu_y)}{\partial \mu_y}$

Seja: $\vec{\alpha} = (\vec{x}_i - \vec{\mu}_y) \Rightarrow \boxed{d\alpha = -d\mu_y}$

$\Rightarrow \frac{\partial f}{\partial \mu} = -\frac{1}{2}\sum_{i=1}^{N} -\frac{\partial (\vec{\alpha})^T [\sigma]^{-1} (\vec{\alpha})}{\partial \alpha} =$

$= \frac{1}{2}\sum_{i=1}^{N} ([\sigma]^{-1} + [\sigma^2]^{-T})\, \alpha =$

$= \frac{1}{2}\sum_{i=1}^{N} \left[\underbrace{([\sigma]^{-1} + [\sigma^{-1}]^T)}_{\beta \text{ (p/ facilitar)}} \cdot (\vec{x}_i - \vec{\mu}_y) \right] = 0 \Rightarrow$

$\Rightarrow \sum_{i=1}^{N} [\beta]\vec{x}_i = \sum_{i=1}^{N} [\beta]\vec{\mu}_y \Rightarrow$

$\Rightarrow \cancel{[\beta]} \sum_{i=1}^{N} \vec{x}_i = \cancel{[\beta]}\, \vec{\mu}_y\, N \Rightarrow$

$$\boxed{\frac{\sum_{i=1}^{N} \vec{x}_i}{N} = \vec{\mu}_y = \hat{\mu}_y}$$