Lecture 16

Monday, November 4, 2024 9:55 AM

Regression

$D = \{ \vec{x}_i, y_i \}_{i=1}^{n}$

$\vec{x}_i \in \mathbb{R}^d$ $\quad$ $y_i \in \mathbb{R}$

$\hat{y}_i = f(\tilde{x}_i) = \tilde{\omega}^T \tilde{x}_i$

linear regression model

$\hat{y}_i = \omega_0 + \omega_1 x_1 + \cdots + \omega_d x_d$

augment $\tilde{D} = \{ \tilde{x}_i, y_i \}$

$\tilde{x}_i = (1 \ \ \vec{x}_i^T)$ ← augmented point

$\tilde{\omega} = (\omega_0 \ \omega_1 \ \cdots \ \omega_d)^T$

Classification:

$D = \{ \vec{x}_i, y_i \}$

$y_i \in$ categorical $= \{ C_1, C_2, \ldots, C_k \}$

k different values

$\hat{y}_i =$ low $\quad$ $y =$ high

$L(y_i, \hat{y}_i) \equiv ?$ loss or error

Do not predict $\hat{y}_i$ directly

Prediction

Sum to 1

$P(y_i = c_1 \mid \vec{x}_i) = \pi_1$

$P(y_i = c_2 \mid \vec{x}_i) = \pi_2$

$\vdots$

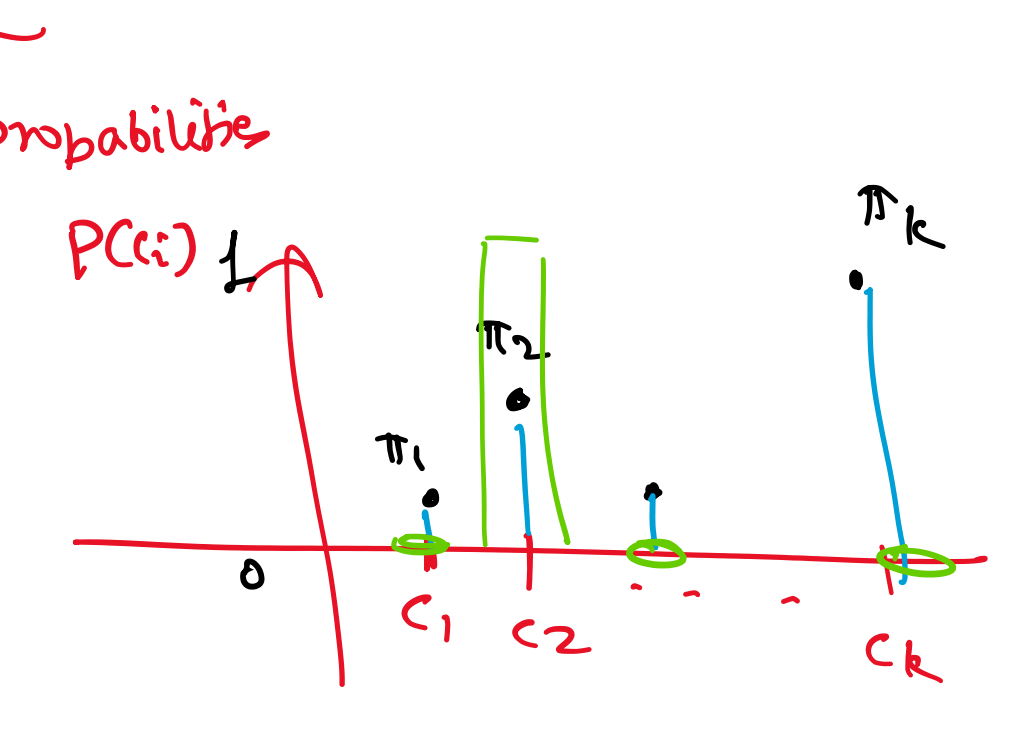
$P(y_i = c_k \mid \vec{x}_i) = \pi_k$

Posterior probabilities

$f_\theta(\vec{x}_i)$

$\theta \equiv$ set of parameters

truth



$\vec{x}_i$ $\quad$ $y_i = C_2$

$\downarrow$

f

$\pi_1, \pi_2, \cdots \pi_k$

Predictions

Sum = 1

$C_1 \ C_2 \ C_3 \cdots C_k$

$0 \ 1 \ 0 \cdots 0$

One-hot encoding

Sum = 1

$y \equiv$ Color

{black, red, orange, blue}
$C_1 \quad C_2 \quad C_3 \quad C_4$

$C_1 = 1000$

$C_2 = 0100$

$C_3 = 0010$

$C_4 = 0001$

$\hat{y} = (0.1 \ 0.4 \ 0.2 \ 0.3)$
$\pi_1 \ \pi_2 \ \pi_3 \ \pi_4$

$L =$ Cross-entropy loss

$\equiv$ NLL $=$ negative log-likelihood

'logistic' regression (binary)

$\vec{x}_i, \ y_i \in \{ C_1 \ C_2 \}$
$\qquad\qquad 1 \quad 0$

binary encoding

one-hot
$\begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$

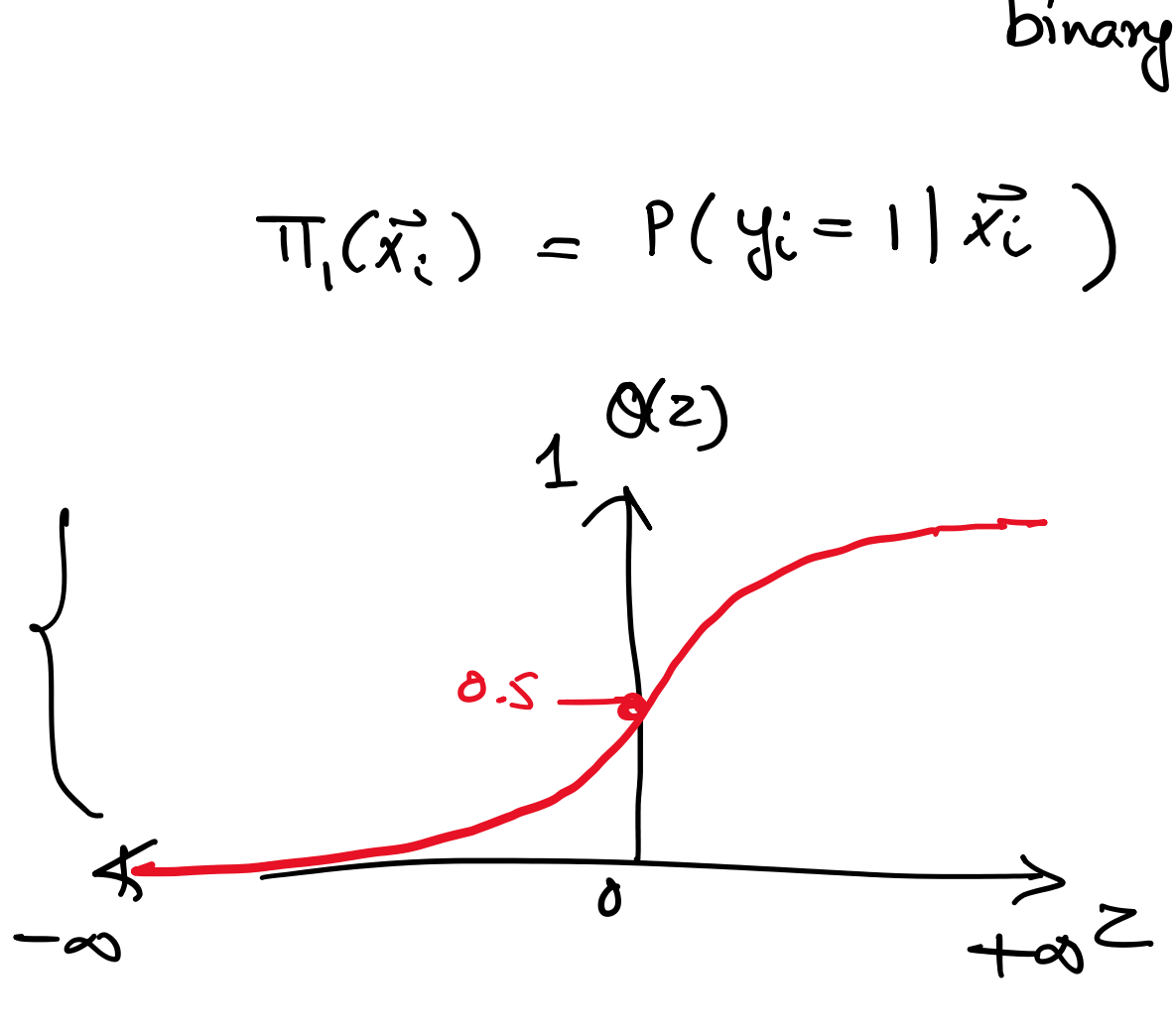
$\pi_1(\vec{x}_i) = P(y_i = 1 \mid \vec{x}_i) = \tilde{\omega}^T \tilde{x}_i$

$\omega_0 + \omega_1 x_{i1} + \omega_2 x_{i2} + \cdots + \omega_d x_{id}$

$[-\infty, \infty]$

$\Downarrow$ Sigmoid function $\equiv \theta$

$[0, 1]$



$\theta(z) = \boxed{\frac{1}{1+e^{-x}}} = \frac{e^x}{1+e^x}$

$\frac{1}{1+e^{-(-\infty)}} \rightarrow \frac{1}{1+\infty} \rightarrow 0$

$\frac{1}{1+e^{-\infty}} \rightarrow 1$

$\frac{1}{1+e^{0}} \rightarrow \frac{1}{2}$

$\pi_1(\tilde{x}_i) = \theta(\tilde{\omega}^T \tilde{x}_i) \equiv$ Prob for C_1

$\pi_2(\tilde{x}_i) = 1 - \theta(\tilde{\omega}^T \tilde{x}_i)$

$1 -$ Prob for C_1

$\pi_1 + \pi_2 = 1$

Constraint

Unknown parameter $\quad \tilde{\omega} = (\omega_0 \ \omega_1 \ \omega_2 \cdots \omega_d)^T$

$d+1$ values

Maximum likelihood $\equiv$ BCE loss

binary cross entropy loss

$\tilde{D} = \{ \tilde{x}_i, y_i \}$

augment

first convert to one-hot vector

$\{ \tilde{x}_i, \vec{y}_i \}$

$\tilde{x}_i \in \mathbb{R}^{d+1}$ $\quad$ $\vec{y}_i \in \{0,1\}^k$

when $k=2$

$y_i = 1$

$y_i = 0$

$y_i =$ black $= (1000)^T \vec{y}_i$

$y_i =$ red $= (0100)^T \vec{y}_i$

$y_i =$ green $= (0010)^T \vec{y}_i$

$y_i =$ blue $= (0001)^T \vec{y}_i$

$y_i =$ Pos — 1

$y_i =$ neg — 0

likelihood

$\max_{\tilde{\omega}} L(D) = P(D \mid \tilde{\omega}) = \prod_{i=1}^{n} P(y_i \mid \tilde{x}_i, \tilde{\omega})$

Prob of observing the data given $\tilde{\omega}$

Prob for $\tilde{x}_i$

assumption all points are IID independent & identically distributed

$P(y_i = 1) = \theta(\tilde{\omega}^T \tilde{x}_i)$ π_1

$P(y_i = 0) = 1 - \theta(\tilde{\omega}^T \tilde{x}_i)$

case statement

$P(y_i \mid x_i) = \theta(\tilde{\omega}^T \tilde{x}_i)^{y_i} (1 - \theta(\tilde{\omega}^T \tilde{x}_i))^{1-y_i}$

$y_i = 1$ $\quad \theta(\tilde{\omega}^T \tilde{x}_i)$

or

$y_i = 0$ $\quad 1 - \theta(\tilde{\omega}^T \tilde{x}_i)$

$\max_{\tilde{\omega}} L(D) = P(D \mid \tilde{\omega}) = \prod_{i=1}^{n} \theta(\tilde{\omega}^T \tilde{x}_i)^{y_i} (1 - \theta(\tilde{\omega}^T \tilde{x}_i))^{1-y_i}$

$\frac{\partial L(D)}{\partial \tilde{\omega}}$ Solve

directly or iteratively (gradient ascent)

$\log L(D) = \sum_{i=1}^{n} \left[\log(\theta(\tilde{\omega}^T \tilde{x}_i)^{y_i}) + \log(1 - \theta(\tilde{\omega}^T \tilde{x}_i))^{1-y_i} \right]$

log-likelihood $\max_{\tilde{\omega}} \log L(D) = \sum_{i=1}^{n} y_i \log(\theta(\tilde{\omega}^T \tilde{x}_i)) + (1-y_i) \log(1 - \theta(\tilde{\omega}^T \tilde{x}_i))$

$\Downarrow$

$\min_{\tilde{\omega}} \text{BCE} = -\sum_{i=1}^{n} y_i \log(\pi_1) + (1-y_i) \log(\pi_2)$

binary cross entropy

$y_i = 1 \quad 0$

$\pi_1 \quad \pi_2$

$0.1 \quad 0.9$

$\nabla_{\tilde{\omega}} = \frac{\partial \text{BCE}}{\partial \tilde{\omega}}$

$\frac{\partial}{\partial \tilde{\omega}} \ y_i \log(\theta(\tilde{\omega}^T \tilde{x}_i))$

log $\quad$ sigmoid $\quad$ linear

$\frac{\partial}{\partial z} \theta(z) = \theta(z)(1 - \theta(z))$

$y_i \cdot \frac{1}{\theta(\tilde{\omega}^T \tilde{x}_i)} \ \theta(\tilde{\omega}^T \tilde{x}_i)(1 - \theta(\tilde{\omega}^T \tilde{x}_i)) \ \tilde{x}_i$

∂log

$y_i (1 - \theta(\tilde{\omega}^T x_i)) x_i$

$\nabla_{\tilde{\omega}} = \frac{\partial \text{BCE}}{\partial \tilde{\omega}} = \sum_{i=1}^{n} (y_i - \theta(\tilde{\omega}^T \tilde{x}_i)) \tilde{x}_i$

true $\quad$ Pred prob

difference in prob

Solve for $\tilde{\omega}$ via gradient descent

$\tilde{\omega}^0 =$ random vector $(d+1)$

for $t = 1 \cdots$ Convergence

$\nabla_{\tilde{\omega}^t} = \sum_{i=1}^{n} (y_i - \theta(\tilde{\omega}^{tT} \tilde{x}_i)) \vec{x}_i$] batch gradient

$\tilde{\omega}^{t+1} = \tilde{\omega}^t - \eta \nabla_{\tilde{\omega}^t}$

step size

Convergence test

$\| \tilde{\omega}^{t+1} - \tilde{\omega}^t \| \leq \epsilon$

threshold

---

gradient descent $\longrightarrow \tilde{\omega}$

$\vec{z}$ new point

$P(y_i = 1 \mid \vec{z}) = \theta(\tilde{\omega}^T \vec{z})$

$P(y_2 = 0 \mid \vec{z}) = 1 - \theta(\tilde{\omega}^T \vec{z})$

if $\theta \geq 0.5$ then $\hat{y}_i = c_1$ (1)

else $\hat{y}_i = c_2$ (0)

assign to the max class

Bayes

$P(y_i = 1 \mid \vec{z}) = f(\vec{x}_i)$

---

full k-way classification

$y_i \in \left\{ \begin{matrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{matrix} \right\} \Rightarrow \vec{y}_i = \left\{ \begin{matrix} 100 \cdots 0 \\ 010 \cdots 0 \\ \cdots 01 \end{matrix} \right\}$

one hot

k

$\tilde{x}_i \rightarrow \pi_1 = P(y_i = c_1 \mid \tilde{x}_i) \ \tilde{\omega}_1$

$\rightarrow \pi_2 = P(y_i = c_2 \mid \tilde{x}_i) \ \tilde{\omega}_2$

$\vdots$

$\pi_k = P(y_i = c_k \mid \tilde{x}_i) \ \tilde{\omega}_k$

$W = \{ \tilde{\omega}_1, \tilde{\omega}_2, \ldots, \tilde{\omega}_k \}$

Unknown parameters

$(d \times 1) \times k$ parameters overall

augmented weight vector $\quad$ # classes

$\pi_1 + \pi_2 + \cdots \pi_k = 1$

dot product

$\pi_1 = \theta(\tilde{\omega}_1^T \tilde{x}_i)$

$\pi_2 = \theta(\tilde{\omega}_2^T \tilde{x}_i)$

$\vdots$

$\pi_k = \theta(\tilde{\omega}_k^T \tilde{x}_i)$

$\sum \pi_i = 1$

$\theta =$ softmax function $(\tilde{\omega}_1 \ \tilde{\omega}_2 \ldots \tilde{\omega}_k)$

$$\pi_j = \theta(\tilde{\omega}_j^T \tilde{x}_i) = \frac{e^{\tilde{\omega}_j^T \tilde{x}_i}}{\sum_{a=1}^{k} e^{\tilde{\omega}_a^T \tilde{x}_i}}$$

$\theta(j) = \frac{\text{exponential (class } j)}{\text{Sum of exponentials for all classes}}$

$\log L(D) = \log P(D \mid W)$

$= \sum_{i=1}^{n} \log P(y_i \mid \tilde{x}_i, W)$

k-cases as one function ?

−NLL

$\min_{W} CE = -\sum_{i=1}^{n} \left[\sum_{j=1}^{k} y_{ij} \log \theta(\tilde{\omega}_j^T x_i) \right]$

softmax

$y_i \log \pi_1 + (1-y_i) \log(1-\pi_1)$
π_2

$\nabla_{\tilde{\omega}_j} = \frac{\partial CE}{\partial \tilde{\omega}_j} = \sum_{i=1}^{n} (y_{ij} - \theta(\tilde{\omega}_j^T x_i)) \tilde{x}_i$

Prob difference

---

1) Initialization

$\tilde{\omega}_j^0 \equiv$ random $(d+1)$ dim vector for $j = 1 \cdots k$

$t = 0$

2) Iterate: gradient descent

$\nabla_{\tilde{\omega}_1^t}, \nabla_{\tilde{\omega}_2^t} \cdots, \nabla_{\tilde{\omega}_k^t}$

$\tilde{\omega}_j^{t+1} = \tilde{\omega}_j^t - \eta \cdot \nabla_{\tilde{\omega}_j^t} \quad \forall j = 1 \cdots k$

3) $t = t+1$

repeat until

$\left(\sum \| \tilde{\omega}_j^{t+1} - \tilde{\omega}_j^t \| \right) \leq \epsilon$

---

test:

$\vec{z} < \begin{matrix} \tilde{\omega}_1^T z \\ \tilde{\omega}_2^T z \\ \vdots \\ \tilde{\omega}_k^T z \end{matrix}$ Softmax $\begin{matrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_k \end{matrix}$ argmax $\hat{y}_i$

Sigmoid $(k=2)$

$\pi_1 = \theta(\tilde{\omega}^T \tilde{x})$

Compute log-odds ratio

$\log\left(\frac{\pi_1}{\pi_2}\right) = \log\left(\frac{\pi_1}{1-\pi_1}\right) = \tilde{\omega}^T \tilde{x}$

logistic function

inverse of sigmoid

$\theta(z) = \frac{1}{1+e^{-z}}$

$\tilde{\omega}^T \tilde{x} \equiv$ log odds ratio of c_1 vs c_2

Softmax $(k>2)$

log odds ratio $= \omega_1$ vs ω_k

$(\tilde{\omega}_1 - \tilde{\omega}_k)^T x_i$

$(\tilde{\omega}_j - \tilde{\omega}_k)^T x_i$

$\equiv$ log odds ratio of c_j vs c_k