Table 1: Dataset

The dataset consists of five samples x_1, x_2, x_3, x_4, x_5. For each sample, we can observe the features A_1, A_2, A_3 and the corresponding response.

Solution:

记 $S = \{x_1, x_2, x_3, x_4, x_5\}$

$$\text{Entropy}(S) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$$

$$= 0.971$$

$$\text{Gain}(S, A_1) = \text{Entropy}(S) - \sum_{v \in \text{Value}(A_1)} \frac{|S_v|}{|S|}\text{Entropy}(S_v)$$

$$= \text{Entropy}(S) - \frac{4}{5}\text{Entropy}(A_1^+) - \frac{1}{5}\text{Entropy}(A_1^-)$$

$$= 0.971 - \frac{4}{5}\times 1 - \frac{1}{5}\times 0$$

$$= 0.171$$

$$\text{Gain}(S, A_2) = \text{Entropy}(S) - \sum_{v \in \text{Value}(A_2)} \frac{|S_v|}{|S|}\text{Entropy}(S_v)$$

$$= 0.971 - \frac{3}{5}\times\text{Entropy}(A_2^+) - \frac{2}{5}\text{Entropy}(A_2^-)$$

$$= 0.971 - \frac{3}{5}\times 0.918 - \frac{2}{5}\times 0$$

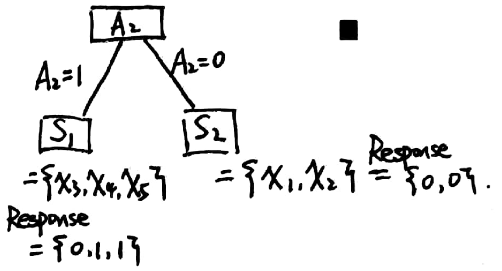
$$= 0.420$$

$$\text{Gain}(S, A_3) = \text{Entropy}(S) - \sum_{v \in \text{Value}(A_3)} \frac{|S_v|}{|S|}\text{Entropy}(S_v)$$

$$= 0.971 - \frac{2}{5}\times\text{Entropy}(A_3^+) - \frac{3}{5}\text{Entropy}(A_3^-)$$

$$= 0.971 - \frac{2}{5}\times 1 - \frac{3}{5}\times 0.918$$

$= 0.020$. 故选择 A_2.



对 $S_1 = \{x_3, x_4, x_5\}$

$$\text{Entropy}(S_1) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$$

$$= 0.918$$

$$\text{Gain}(S_1, A_1) = \text{Entropy}(S_1) - \sum_{v \in \text{Value}(A_1)} \frac{|S_v|}{|S_1|}\text{Entropy}(S_v)$$

$$= 0.918 - \frac{1}{3}\times 0 - \frac{2}{3}\times 0$$

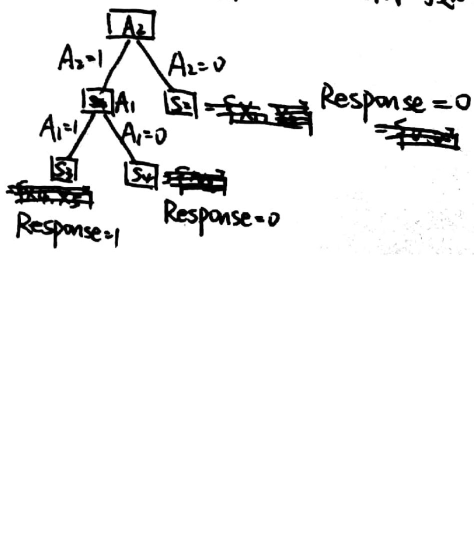
$$= 0.918$$

$$\text{Gain}(S_1, A_3) = \text{Entropy}(S_1) - \sum_{v \in \text{Value}(A_3)} \frac{|S_v|}{|S_1|}\text{Entropy}(S_v)$$

$$= 0.918 - \frac{1}{3}\times 0 - \frac{2}{3}\times 1$$

$$= 0.251$$

故选择 A_1，最终的决策树如下：



$$g(x) = -\sum_{i=1}^{n} H_i \log(f_i(x)),$$

where $H \in \mathbb{R}^n$ is a one-hot vector.

Solution: ■

1、 $$\frac{\partial f_i(x)}{\partial x_i} = \frac{\exp(x_i)\sum_{k=1}^{n}\exp(x_k) - (\exp(x_i))^2}{(\sum_{k=1}^{n}\exp(x_k))^2}$$

$$= \frac{\exp(x_i)}{\sum_{k=1}^{n}\exp(x_k)} - \left(\frac{\exp(x_i)}{\sum_{k=1}^{n}\exp(x_k)}\right)^2$$

$$= f_i(x) - (f_i(x))^2$$

$$\frac{\partial f_i(x)}{\partial x_j} = \frac{-\exp(x_i)\exp(x_j)}{(\sum_{k=1}^{n}\exp(x_k))^2}$$

$$= -f_i(x) f_j(x)$$

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & & & \\ \vdots & & & \\ \frac{\partial f_n}{\partial x_1} & \cdots & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

$$= \begin{bmatrix} f_1 - f_1^2 & -f_1 f_2 & \cdots & -f_1 f_n \\ -f_1 f_2 & f_2 - f_2^2 & & \\ \vdots & & \ddots & \\ -f_1 f_n & & & f_n - f_n^2 \end{bmatrix}$$

$$Df(x) = (\nabla f(x))^T = \nabla f(x)$$

2. $$f_i(x-c) = \frac{\exp(x_i - c)}{\sum_{k=1}^{n}\exp(x_k - c)}$$

$$= \frac{\exp(x_i)/\exp(c)}{\sum_{k=1}^{n}\exp(x_k)/\exp(c)}$$

$$= \frac{\exp(x_i)}{\sum_{k=1}^{n}\exp(x_k)} = f_i(x)$$

$\because f(x) = f(x-c)$.

$x_i - c = x_i - \max_i\{x_i\} \leq 0$

在进行运算时(求导等)需要控制指数的大小,
防止(如 f_i^2 或 $f_i f_j$)溢出.

$\exp\{x_i - c\} \leq 1$ 则可以控制大小。

3、 $$\frac{\partial g(x)}{\partial x_j} = -\sum_{i=1}^{n} H_i \frac{\partial \log(f_i(x))}{\partial x_j}$$

$$= -\sum_{i=1}^{n} H_i \frac{\partial \log(f_i(x))}{\partial f_i(x)} \cdot \frac{\partial f_i(x)}{\partial x_j}$$

$$= -\sum_{i=1}^{n} H_i \frac{1}{f_i(x)} \cdot \frac{\partial f_i(x)}{\partial x_j}$$

$$= -\sum_{i=1, i\neq j}^{n} H_i \frac{-f_i(x) f_j(x)}{f_i(x)} - H_j \frac{f_j(x) - f_j^2(x)}{f_j(x)}$$

$$= \sum_{i=1, i\neq j}^{n} H_i f_j(x) + H_j f_j(x) - H_j$$

$$= f_j(x) - H_j$$

(a) The convolutinal layer parameters are denoted as "conv⟨filter size⟩-⟨number of filters⟩".

(b) The fully connected layer parameters are denoted as "FC⟨number of neurons⟩".

(c) The window size of pooling layers is 3.

(d) The stride of convolutinal layers is 1.

(e) The stride of pooling layers is 3.

(f) There is no padding in both convolutional and pooling layers.

(g) For convenience, we assume that there is no activation function and bias.

Suppose that the input is a **386** $\times$ **386 RGB** image. Please derive the size of all feature maps and the number of parameters.

conv3-64	max pool	conv3-256	conv1-512	max pool	FC-2048	FC-1000

Table 2: The architecture of convolutional neural network

Solution:

1. $\frac{\partial f_1(x)}{\partial x_i} = \frac{1}{n}$.

$$\nabla f_1(x) = \left.\begin{pmatrix} \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{pmatrix}\right\} n个.$$

3.　2.

con (a) 先考虑二维 $\{x_1, x_2\}$ 的情况

$$f_2(x) = \max\{x_1, x_2\}$$
$$= \frac{1}{2}(x_1 + x_2 + |x_1 - x_2|)$$

m

在 $x_1 = x_2$ 处 $|x_1 - x_2|$ 不可微

$x_1 \neq x_2$ 处 $x_1 + x_2$, $|x_1 - x_2|$ 均可微,

故 $f_2(x)$ 可微.

考虑多维情况

设 x_i, x_j 为 $x_1, x_2, \cdots x_n$ 中最大的两个数

$$f_2(x) = \max\{x_1, x_2, \cdots x_n\} = \max\{x_i, x_j\}$$
$$= \frac{1}{2}(x_i + x_j + |x_i - x_j|)$$

同上讨论知在 $x_i = x_j$ 时 $f_2(x)$ 不可微

$x_i \neq x_j$ 处 $f_2(x)$ 可微.

故可微点组合的集合 C 为

$$C = \{(x_1, x_2 \cdots x_n) \mid x_j = \max_i\{x_i\}, x_k = \max_{i, i \neq j}\{x_i\}, x_j \neq x_k\}$$

(b)

令 $x_j = \max_i\{x_i\}$

选取 $d(x)$ 为

$$d(x)_i = \begin{cases} 0, & \text{其他} \\ 1, & i = j \end{cases}$$

则

$$\cancel{f_2(y)}\ f_2(x) + \langle d(x), y - x\rangle$$
$$= f_2(x) + y_j - x_j$$
$$= x_j + y_j - x_j = y_j \leq \max_i\{y_i\} = f_2(y).$$

3.

	feature maps	parameters
conv3-64	$64\times(386-3+1)\times(386-3+1)$ $=9437184$	9×64 $=576$
max-pool	$64\times\frac{384}{3}\times\frac{384}{3}$ $=1048576$	无
conv3-256	$256\times64\times(\frac{384}{3}-3+1)\times(\frac{384}{3}-3+1)$ $=260112384$	9×256 $=2304$
conv1-512	$512\times256\times64\times(126-1+1)(126-1+1)$ $=133177540608$	1×512 $=512$
max-pool	$512\times256\times64\times\frac{126}{3}\times\frac{126}{3}$ $=14797504512$	无
FC-2048	2048	$14797504512\times2048\times3$ $=90915867721728$
FC-1000	1000	2048×1000 $=2048000$

经过第一个FC后的特征已当不区分RGB三个通道，故无需设计×3的权重参数

王

Exercise 4: Matrix Calculus 20pts

Let $L = f(h(Ax+b))$, where $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, and $f : \mathbb{R}^m \to \mathbb{R}$. Define $z = Ax + b \in \mathbb{R}^m$ and $w = h(z) = (\sigma(z_1), \dots, \sigma(z_n))^\top$, where z_i is the i^{th} component of z and

$$\sigma(z_i) = \frac{1}{1+\exp(-z_i)}.$$

Assume $\nabla_w f$ is known.

1. Please derive $\nabla_x L$.

2. Please derive

$$\nabla_A L = \begin{bmatrix} \frac{\partial L}{\partial A_{11}} & \cdots & \frac{\partial L}{\partial A_{1j}} & \cdots & \frac{\partial L}{\partial A_{1n}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{\partial L}{\partial A_{i1}} & \cdots & \frac{\partial L}{\partial A_{ij}} & \cdots & \frac{\partial L}{\partial A_{in}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{\partial L}{\partial A_{m1}} & \cdots & \frac{\partial L}{\partial A_{mj}} & \cdots & \frac{\partial L}{\partial A_{mn}} \end{bmatrix},$$

where $A_{i,j}$ is the entry in the i^{th} row, j^{th} column of the matrix A.

Solution: ■

1、

$$\nabla_x L = \nabla_x z \, \nabla_z w \, \nabla_w f$$

$$\nabla_x z = A^T$$

$$(\nabla_z w)_{ij} = \begin{cases} \frac{\exp(-z_j)}{(1+\exp(-z_j))^2}, & i=j \\ 0, & i \neq j \end{cases}$$

PB17000215
张博文

$$\nabla_x L = A^T \begin{pmatrix} \frac{\exp(-z_1)}{(1+\exp(-z_1))^2} & & \\ & \ddots & \\ & & \frac{\exp(-z_m)}{(1+\exp(-z_m))^2} \end{pmatrix} \nabla_w f$$

2、

$$\frac{\partial L}{\partial A_{ij}} = \cancel{\partial z}\frac{\partial f}{\partial w_i} \cdot \frac{\partial w_i}{\partial z_i} \cdot \frac{\partial z_i}{\partial A_{ij}}$$

$$= (\nabla_w f)_i \, \frac{\exp(-z_i)}{(1+\exp(-z_i))^2} x_j$$