Multi-class classification

f() $\longrightarrow$ {Cat, Dog, Mouse,...}

$$y = f(\mathbf{x}), \quad \text{Input: } \mathbf{x} \in \mathbb{R}^n \longrightarrow \text{Output: } y \in \{1, 2, \ldots, C\}$$

Ordering irrelevant!

1: Cat, 2: Dog, 3: Mouse

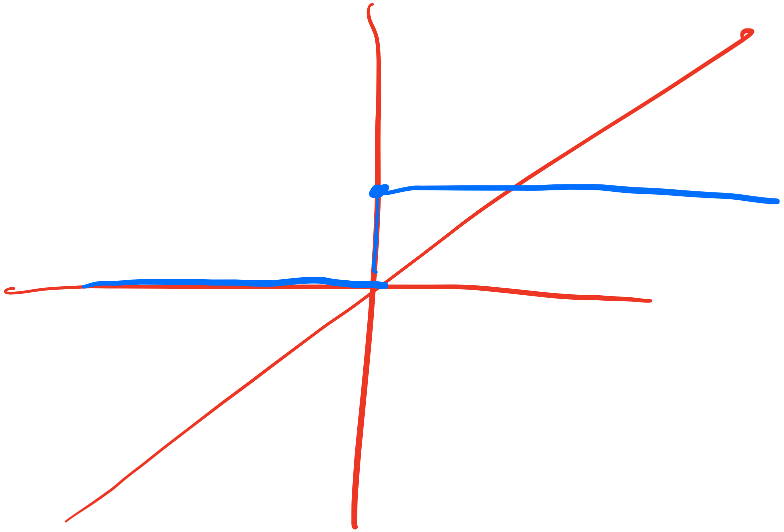
1: Dog, 2: Mouse, 3: Cat

Multi-class prediction functions

Binary thresholding

$$f(\mathbf{x}) = \mathbb{I}(\underbrace{\mathbf{x}^T\mathbf{w}}_{\text{linear function}} \geq 0) \rightarrow \{0, 1\}$$

Indicator



Multi-class thresholding

$$f(\mathbf{x}) = \underset{c \in \{1...C\}}{\text{argmax}}\ \mathbf{x}^T\mathbf{w}_c$$

Make C different prediction

Multi-class prediction functions

Multi-class thresholding

$$f(\mathbf{x}) = \underbrace{\underset{c \in \{1 \ldots C\}}{\text{argmax}}}_{C \text{ linear func.}} \ \mathbf{x}^T \mathbf{w}_c$$

For convenience, we can also define a matrix that contains all C parameter vectors:

$$\mathbf{W} = \begin{bmatrix} \mathbf{w}_{\boxed{1}}^T \\ \mathbf{w}_{\boxed{2}}^T \\ \vdots \\ \mathbf{w}_{\boxed{C}}^T \end{bmatrix} = C \underbrace{\overbrace{\begin{bmatrix} W_{11} & W_{12} & \dots & W_{1d} \\ W_{21} & W_{22} & \dots & W_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ W_{C1} & W_{C2} & \dots & W_{Cd} \end{bmatrix}}^{d}}_{C \times d}$$

$C \times d$ total params

\# num of measurement

\# of classes

With this notation, our prediction function becomes:

$$f(\mathbf{x}) = \underset{c \in \{1 \ldots C\}}{\text{argmax}} \ (\mathbf{x}^T \mathbf{W}^T)_c, \quad \mathbf{W} \in \mathbb{R}^{C \times d}$$

x: $d \times 1$ column

x^T: $1 \times d$ row vector

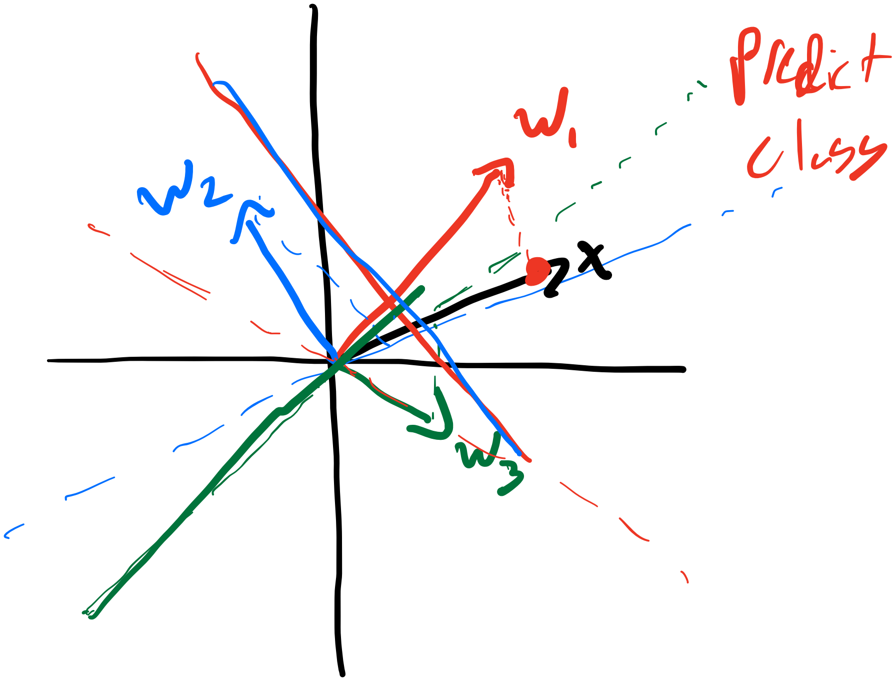
W: $C \times d$ matrix, W^T: $d \times C$ matrix

$(1 \times d)(d \times C) \rightarrow (1 \times C)$ vector

$(x^T W^T)_c = \boxed{x^T W_c}$ scalar

Multi-class decision boundaries

$$f(\mathbf{x}) = \underset{c\in\{0,1\}}{\text{argmax}}\ (\mathbf{x}^T\mathbf{W}^T)_c = \mathbb{I}(\mathbf{x}^T\mathbf{w}_1 - \mathbf{x}^T\mathbf{w}_0 \geq 0)$$



$C = 2$

$x^T w_0 \geq x^T w_1$? if false 1, if true 0

$x^T w_0 - x^T w_1 \geq 0$

↓ 2 parameter

$x^T \underbrace{(w_0 - w_1)}_{\text{def. } w} \geq 0$

$= x^T w \geq 0$

Logistic ↙ Reg

Categorical distribution

C total probs

Prob. of outcome c

$$p(y = c) = q_c, \quad y \in \{1 \ldots C\}$$

$$\mathbf{q} \in \mathbb{R}^C \quad q_c \geq 0 \; \forall c \in \{1 \ldots C\} \quad \sum_{c=1}^{C} q_c = 1$$

Sum to 1

Fair die: $\{1, 2, 3, 4, 5, 6\}$

$p(y=c) = \frac{1}{6} \; \frac{1}{6} \; \frac{1}{6} \; \frac{1}{6} \; \frac{1}{6} \; \frac{1}{6}$

A probabilistic model for multi-class classification

Model $P(y \mid x)$ = categorical

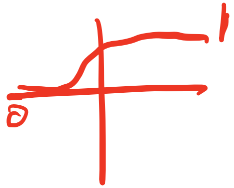
$$\underbrace{y_i}_{\text{output / label}} \sim \text{Categorical}(\boxed{\mathbf{q} = ?}), \quad \mathbf{q} = \mathbf{x}_i^T\mathbf{W}^T ? \longrightarrow \text{vector length } C$$

$$\mathbf{x}^T\mathbf{W}^T \in \mathbb{R}^C, \quad q_c \ngeq 0 \; \forall c \in \{1 \ldots C\}, \quad \sum_{c=1}^{C} q_c \neq 1$$

$q = ?$

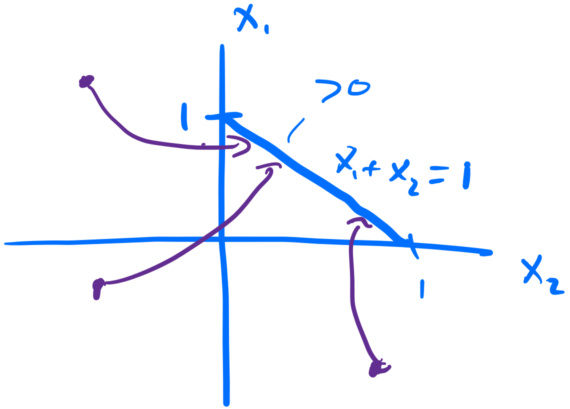
Need $f(\mathbf{x}) : \mathbb{R}^C \longrightarrow [0, \infty)^C, \quad \sum_{i=1}^{C} f(\mathbf{x})_c = 1$

logistic Reg.



$$x^T W \in \mathbb{R} \xrightarrow{\sigma} [0, 1]$$

sigmoid func.



Categorical distribution

$$p(y=c) = q_c, \quad y \in \{1 \ldots C\}$$

total # of classes or categories

Bern.

$P(y=1) = q$

$$P(y) = (q)^y (1-q)^{(1-y)}$$

$$\mathbf{q} \in \mathbb{R}^C \quad q_c \geq 0 \; \forall c \in \{1 \ldots C\} \quad \sum_{c=1}^{C} q_c = 1$$

index of the summation

$$p(y) = \prod_{c=1}^{C} \boxed{q_c^{\mathbb{I}(y=c)}} = \boxed{q_y}$$

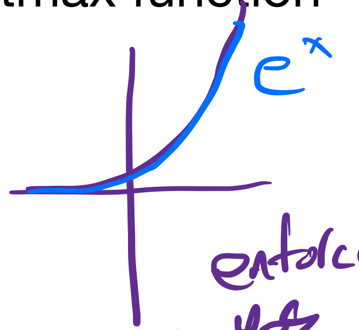
$$\begin{cases} q_1 & \text{if } y=1 \\ q_2 & \text{if } =2 \\ \vdots & \end{cases}$$

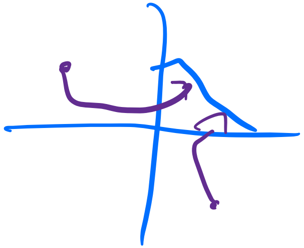
$$\log p(y) = \log \prod q_c^{\mathbb{I}(y=c)} = \sum_{c=1}^{C} \mathbb{I}(y=c) \log q_c$$

Softmax function

$\mathbb{R}^C \to \mathbb{R}^C$

$$\mathrm{softmax}(\mathbf{x})_{c} = \frac{e^{x_c}}{\boxed{\sum_{j=1}^{C} e^{x_j}}}$$





enforce 1) pos.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_c \end{bmatrix} \longrightarrow \begin{bmatrix} e^{x_1} \\ e^{x_2} \\ \vdots \\ e^{x_c} \end{bmatrix}$$

$$\mathrm{softmax}(\mathbf{x}) = \begin{array}{c} \begin{bmatrix} \boxed{\frac{e^{x_1}}{\sum_{j=1}^{C} e^{x_j}}} \\ \frac{e^{x_2}}{\sum_{j=1}^{C} e^{x_j}} \\ \vdots \\ \frac{e^{x_C}}{\sum_{j=1}^{C} e^{x_j}} \end{bmatrix} \end{array} \begin{array}{c} 1 \\ 2 \\ \vdots \\ C \end{array}$$

2) enforce sum to 1

$$T = \sum_{j=1}^{C} e^{x_j}$$

$$\frac{1}{T} \sum_{j=1}^{C} e^{x_j}$$

$$\begin{bmatrix} e^{x_1} \\ e^{x_2} \\ \vdots \\ e^{x_C} \end{bmatrix} \longrightarrow \begin{bmatrix} \frac{e^{x_1}}{\sum_{j=1}^{C} e^{x_j}} \\ \vdots \end{bmatrix}$$

Softmax function

$$\mathrm{softmax}(\mathbf{x})_c = \frac{e^{x_c}}{\sum_{j=1}^{C} e^{x_j}}$$

$$\underbrace{\sum_{i=1}^{C} \frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}} = \frac{\sum_{i=1}^{C} e^{x_i}}{\sum_{j=1}^{C} e^{x_j}}} = 1$$

$$\boxed{\operatorname*{argmax}_{c\in\{1,\ldots,C\}} \mathbf{x}_c = \operatorname*{argmax}_{c\in\{1,\ldots,C\}} \mathrm{softmax}(\mathbf{x})_c}$$

A probabilistic model for multi-class classification: Multinomial Logistic regression

$$\boxed{y_i \sim \textbf{Categorical}\big(\overbrace{\text{softmax}(\mathbf{x}^T\mathbf{W}^T)}^{q}\big)}$$

linear func. c

$$p(y_i = c \mid \mathbf{x}, \mathbf{W}) = \text{softmax}(\mathbf{x}^T\mathbf{W}^T)_c = \frac{e^{\mathbf{x}^T\mathbf{w}_c}}{\sum_{j=1}^{C} e^{\mathbf{x}^T\mathbf{w}_j}}$$

input parameters

index c

← summed

all linear func.

exp.

Maximum likelihood estimation for Multinomial Logistic regression

Model

$$y_i \sim \textbf{Categorical}(\text{softmax}(\mathbf{x}^T\mathbf{W}))$$

$$p(y_i = c \mid \mathbf{x}, \mathbf{W}) = \text{softmax}(\mathbf{x}^T\mathbf{W})_c = \frac{e^{\mathbf{x}^T\mathbf{w}_c}}{\sum_{j=1}^{C} e^{\mathbf{x}^T\mathbf{w}_j}}$$

Loss

$$\mathbf{Loss}(\mathbf{W}) = \mathbf{NLL}(\mathbf{W}, \mathbf{X}, \mathbf{y}) = -\sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i, \mathbf{W})$$

output Input

$$\mathbf{NLL}(\mathbf{W}, \mathbf{X}, \mathbf{y}) = -\sum_{i=1}^{N} \log\left[\frac{e^{x^T w_{y_i}}}{\sum_{j=1}^{C} e^{x^T w_j}}\right]$$

$$= -\sum_{i=1}^{N} \cancel{\log e}^{\,x^T w_{y_i}} - \underbrace{\log \sum_{j=1}^{C} e^{x^T w_j}}$$

$$= -\sum_{i=1}^{N} x^T w_{y_i} - \log \sum_{j=1}^{C} e^{x^T w_j}$$