CSC 311: Introduction to Machine Learning

Lecture 5 - Linear Models III

Michael Zhang Chandra Gummaluru

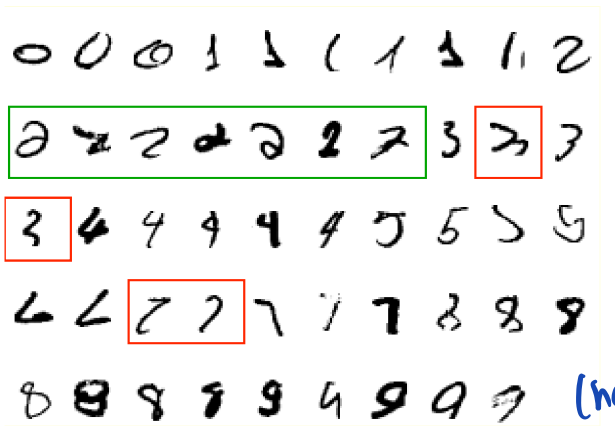
University of Toronto, Winter 2023

Outline

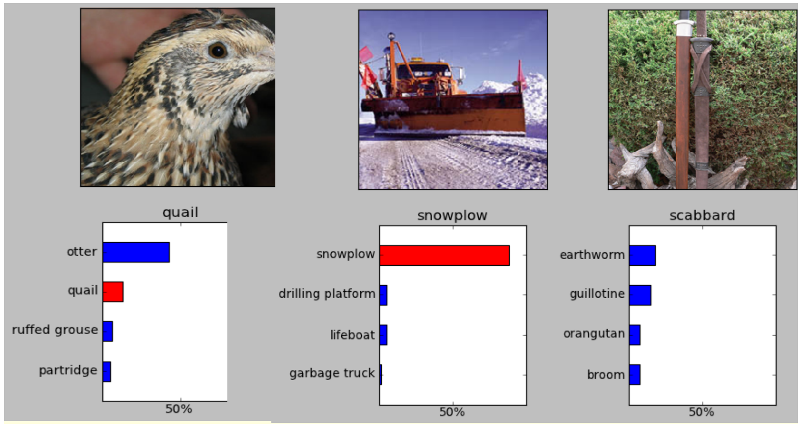
1. Softmax Regression
2. Tracking Model Performance
3. Limits of Linear Classification
4. Midterm Review
5. Introducing Neural Networks
6. Expressivity of a Neural Network

Multi-class Classification

Task is to predict a discrete(> 2)-valued target.



(hardest examples)



MNIST

FashionMNIST

CIFAR-10

-100

ImageNet (1000 classes)

Targets in Multi-class Classification

- Targets form a discrete set $\{1, \ldots, K\}$.
- Represent targets as one-hot vectors or one-of-K encoding:

$$\mathbf{t} = \underbrace{(0, \ldots, 0, 1, 0, \ldots, 0)}_{\text{entry } k \text{ is } 1} \in \mathbb{R}^K$$

output space

$N \times D$

K-dim

$D \times (K+1)$

bias

Linear Function of Inputs

Vectorized form:

$$\mathbf{z} = \mathbf{W}\mathbf{x} + \mathbf{b} \text{ or}$$

$$\mathbf{z} = \mathbf{W}\mathbf{x} \text{ with dummy } x_0 = 1$$

K : number of classes

D : number of features

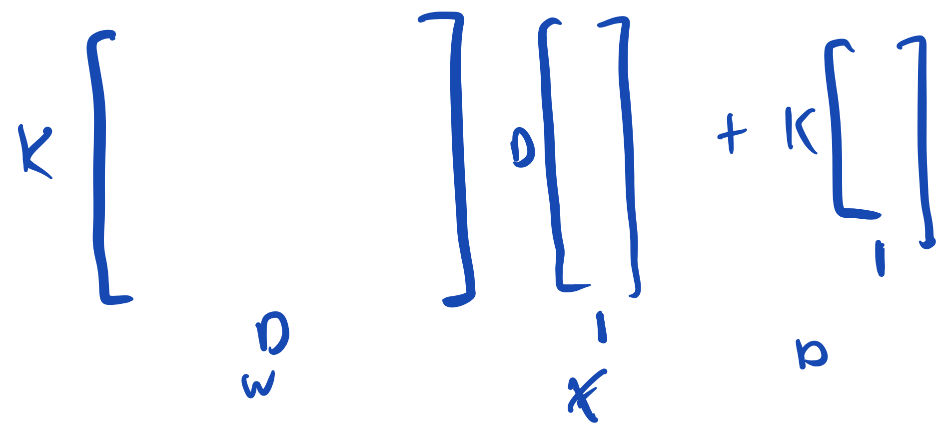
X.T

Non-vectorized form:

$$z_k = \sum_{j=1}^{D} w_{kj} x_j + b_k \quad \text{for} \quad k = 1, 2, ..., K$$

output x input

- $\mathbf{W}$: K x D matrix.
- $\mathbf{x}$: D x 1 vector.
- $\mathbf{b}$: K x 1 vector.
- $\mathbf{z}$: K x 1 vector.



K×1 + K×1 = K×1

Generating a Prediction

Interpret z_k as how much the model prefers the k-th prediction.

$$y_i = \begin{cases} 1, & \text{if } i = \arg\max_k z_k \\ 0, & \text{otherwise} \end{cases}$$

(similar to 0-1)

not easily differentiable

How does the $K = 2$ case relate to the binary linear classifiers?

Softmax Regression

- Soften the predictions for optimization.
- A natural activation function is the softmax function, a generalization of the logistic function: np.exp

$z = [-2, 0, 3, \dots]$

$e^{-2}\ e^{0}\ e^{3}, \dots$

$$y_k = \text{softmax}(z_1, \dots, z_K)_k = \frac{e^{z_k}}{\sum_{k'} e^{z_{k'}}}$$

normalize

- Inputs z_k are called the logits.
- Interpret outputs as probabilities. y
- If z_k is much larger than the others, then $\text{softmax}(\mathbf{z})_k \approx 1$ and it behaves like argmax.

$z_1 = w_1 x$

$z_2 = w_2 x$

What does the $K = 2$ case look like?

$$y_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2}} = \frac{e^{w_1 x}}{e^{w_1 x} + e^{w_2 x}}$$

Cross-Entropy as Loss Function

$\sigma(u) = \frac{1}{1+e^{-u}}$

correct pred
loss is 0
predict 0.99
penalty >>
predict 0.8

$$\frac{1}{1+e^{w_2 x - w_1 x}}$$

Use cross-entropy as the loss function.

$$\mathcal{L}_{\mathrm{CE}}(\mathbf{y}, \mathbf{t}) = -\sum_{k=1}^{K} t_k \log y_k = -\mathbf{t}^\top (\log \mathbf{y}),$$

$$\frac{1}{1+e^{(w_2 - w_1)x}}$$

Single
matrix
W in
logistic
regression

where the log is applied element-wise.

Often use a combined softmax-cross-entropy function.

↳ information theory meaning

resume: 12:20

Gradient Descent Updates for Softmax Regression

Softmax Regression:

$$
\begin{aligned}
\mathbf{z} &= \mathbf{W}\mathbf{x} \\
\mathbf{y} &= \operatorname{softmax}(\mathbf{z}) \\
\mathcal{L}_{\mathrm{CE}} &= -\mathbf{t}^{\top}(\log \mathbf{y})
\end{aligned}
$$

Gradient Descent Updates:

$$
\begin{aligned}
\frac{\partial \mathcal{L}_{\mathrm{CE}}}{\partial \mathbf{w}_k} = \frac{\partial \mathcal{L}_{\mathrm{CE}}}{\partial z_k} \cdot \frac{\partial z_k}{\partial \mathbf{w}_k} &= (y_k - t_k) \cdot \mathbf{x} \\
\mathbf{w}_k \leftarrow \mathbf{w}_k - \alpha \frac{1}{N} \sum_{i=1}^{N} (y_k^{(i)} - t_k^{(i)}) \mathbf{x}^{(i)}
\end{aligned}
$$

$$(X^TX + \lambda I)\, w^* = X^T t$$ direct solution

matrix invertible

problem $\underbrace{\|Xw - t\|_2^2}_{\text{(linear reg)}} + \lambda \|w\|_2^2$

$$\|w\|_2^2 = w_1^2 + w_2^2 + \dots + w_d^2$$
$$= w^T w = w \cdot w$$

$$\nabla_w \|w\|_2^2 = 2w$$

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

$$w^T w$$
$$= w^T I w$$

$$\begin{bmatrix} \frac{\partial f}{\partial w_1} \\ \frac{\partial f}{\partial w_2} \\ \vdots \end{bmatrix} = \begin{bmatrix} 2w_1 \\ 2w_2 \\ \vdots \\ 2w_d \end{bmatrix}$$

$$= 2w$$

$f(x) = \underbrace{x^T A x}_{}$

quadratic form

Multivariate Gaussians

$$\nabla_w\, w^T I w$$
$$= 2Iw$$
$$= 2w$$

$$\propto e^{-\frac{x^2}{2}}$$

$$x : \mathbb{R}^n$$
$$f \downarrow$$
$$\mathbb{R}$$

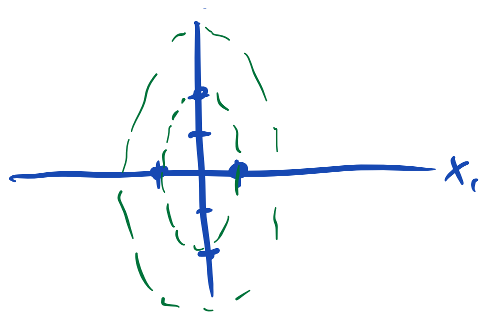
$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= a_{11} x_1^2 + a_{12} x_1 x_2$$
$$+ a_{21} x_2 x_1 + a_{22} x_2^2 = \sum_i \sum_j A_{ij} x_i x_j$$

x_2

$$x^T \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} x$$

$$= 4x_1^2 + x_2^2 = r$$



$r = 4$

$x_1 = \pm 1$

$x_2 = \pm 2$

$$\begin{bmatrix} x \end{bmatrix}^T \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & & & \\ \vdots & & & \\ & & a_{ij} & \\ & & & a_{nn} \end{bmatrix} \begin{bmatrix} x \end{bmatrix}$$

i_{th} i_{th}

$$\frac{\partial f}{\partial x_i} = \frac{d\left[\sum_{j \neq i} a_{ij} x_i x_j + \sum_{j \neq i} a_{ji} x_i x_j + a_{ii} x_i^2\right]}{dx_i}$$

$$= \sum_{j \neq i} a_{ij} x_j + \sum_{j \neq i} a_{ji} x_j + 2a_{ii} x_i$$

$$= \sum_{j=1}^{n} a_{ij} x_j + \sum_{j=1}^{n} a_{ji} x_j$$

$2Ax$ symmetric

$(A + A^T)x$ not sym.