Multinomial Logistic Regression

CSC 461: Machine Learning

Fall 2020

Prof. Marco Alvarez
University of Rhode Island

Logistic regression

- Binary classification
  - ✓ uses a **logistic function** (type of sigmoid function)
  - ✓ models **probability** of output in terms of input
- What if we want k > 2 classes?
  - ✓ can try one-vs-all
    - ✓ learn a binary classifier per class; relabel training data with samples of that class as positive and all other as negatives; predict using the highest score from all classifiers
  - ✓ can try one-vs-one
    - ✓ learn k (k-1) / 2 binary classifiers; each learns to distinguish between two classes; predict using a voting scheme

Issues with OvA or OvO

- Class imbalance
- Scale of scores may differ from classifier to classifier
- Computational cost (both train and predict)

MNIST

- The MNIST database is a large database of handwritten digits
  - ✓ contains 60,000 training images and 10,000 testing images
  - ✓ convolutional neural networks, manages to get an error rate of 0.23%
  - ✓ original paper reports an error rate of 0.8% with SVMs

http://yann.lecun.com/exdb/mnist/

https://en.wikipedia.org/wiki/MNIST_database

Basics of multiclass classification

- Data instance
  - ✓ in general, $x \in \mathcal{X}, \mathcal{X} = \mathbb{R}^d$
  - ✓ $y \in \mathcal{Y}, \mathcal{Y} = \{1,2,\ldots,C\}$
- Hypothesis
  - ✓ each hypothesis **g** is a classifier

$$g : \mathcal{X} \mapsto \mathcal{Y}, g \in \mathcal{H}$$

From binary to k classes

- Binary logistic regression:

$$P(y = +1 \mid \mathbf{x}; \mathbf{w}) = \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}}} = \frac{e^{\mathbf{w}^T\mathbf{x}}}{e^{\mathbf{w}^T\mathbf{x}} + 1}$$

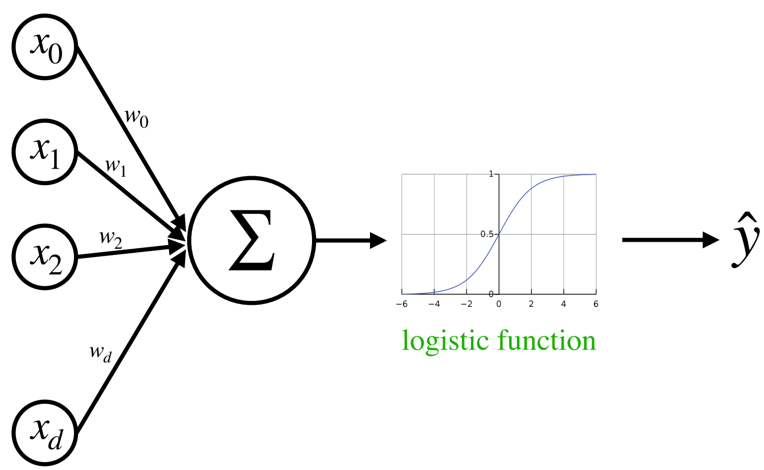
- Can be extended to:

$$P(y = c \mid \mathbf{x}; \mathbf{W}) = \frac{e^{\mathbf{w}_c\mathbf{x}}}{\sum_{k=1}^{C} e^{\mathbf{w}_k\mathbf{x}}}$$

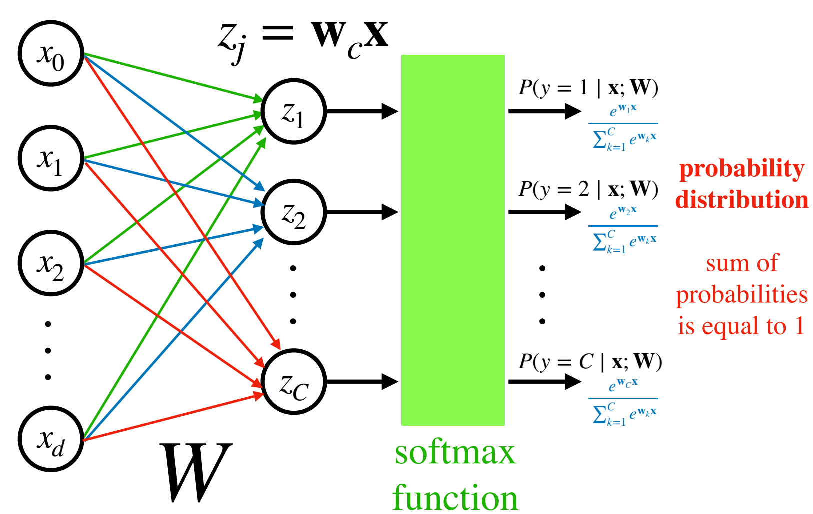
$\mathbf{w}_c$ is the row vector c from $\mathbf{W}$

$W_{C \times d+1}$ is a matrix where every row is a “class” weight vector

Logistic regression (binary classification)



Multinomial logistic regression



Example

- What is the value of softmax($\mathbf{z}$), given that $\mathbf{z}^T = [-10, 10, 5, 4.3, 7]$?

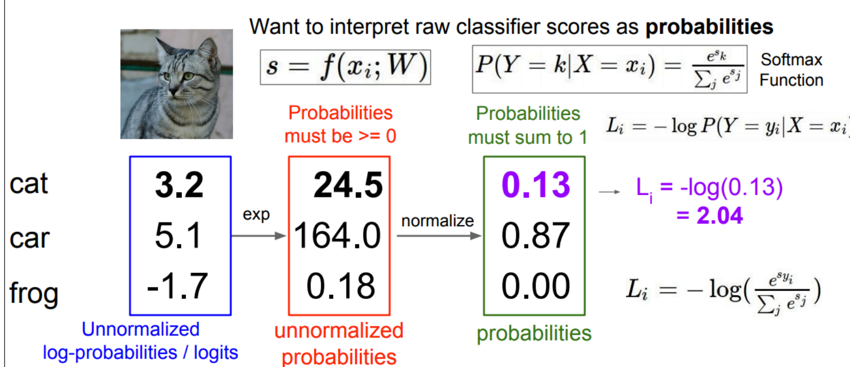
Multinomial logistic regression

- Use the **softmax function** for activation
- Predict the label with the highest probability score (forward pass)

$$\hat{y} = \arg\max_{c} P(y = c \mid \mathbf{x}; \mathbf{W})$$

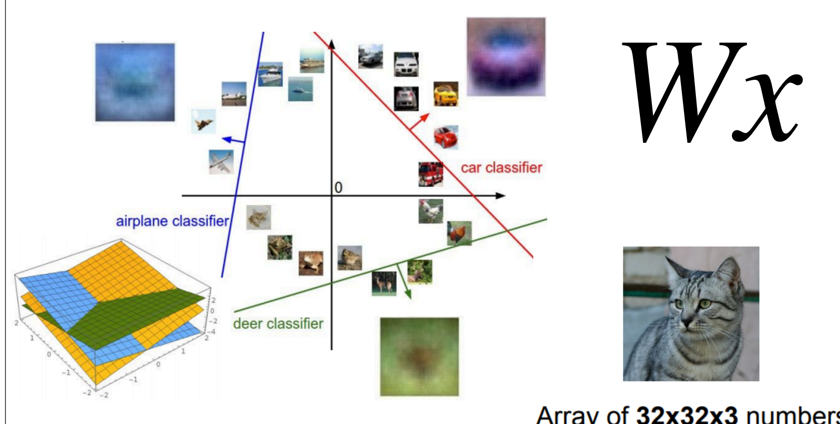
- How to learn the weights?
  - ✓ need to define a **Loss Function** … then apply **gradient descent**
  - ✓ loss function can be derived using **MLE** (similar to binary logistic regression)

Softmax and the loss function



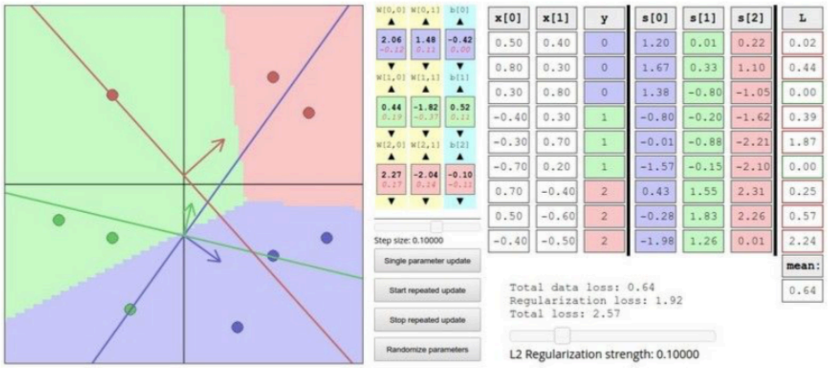
http://vision.stanford.edu/teaching/cs231n/slides/2019/cs231n_2019_lecture03.pdf

Geometric interpretation



http://vision.stanford.edu/teaching/cs231n/slides/2019/cs231n_2019_lecture02.pdf

Interactive web demo



http://vision.stanford.edu/teaching/cs231n-demos/linear-classify/