Data Mining

Lecture 15

Ananya Jana
CS360

Fall 2024

Convolutional neural network (CNN)

- **Background: Computer Vision**
  - Image Classification
  - ILSVRC 2010 - 2016
  - Traditional Feature Extraction Methods
  - Convolution as Feature Extraction
- **Convolutional Neural Networks (CNNs)**
  - Learning Feature Abstractions
  - Common CNN Layers:
    - Convolutional Layer
    - Max-Pooling Layer
    - Fully-connected Layer (w/tensor input)
    - Softmax Layer
    - ReLU Layer
  - Background: Subgradient
  - Architecture: LeNet
  - Architecture: AlexNet
- **Training a CNN**
  - SGD for CNNs
  - Backpropagation for CNNs

Slides taken from https://www.cs.cmu.edu/~mgormley/courses/10601-s17/slides/lecture21-cnn.pdf

Examination Thursday (24th Oct' 24)

Full Marks: 60
Mix of Short Answer Type, Long Answer Type and MCQs.
You can bring your calculator.

Convolutional neural network (CNN)

Background

A Recipe for Machine Learning

1. Given training data:

$$\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N}$$

2. Choose each of these:
   - Decision function

$$\hat{\boldsymbol{y}} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$$

   - Loss function

$$\ell(\hat{\boldsymbol{y}}, \boldsymbol{y}_i) \in \mathbb{R}$$

3. Define goal:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \boldsymbol{y}_i)$$

4. Train with SGD:
(take small steps opposite the gradient)

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta_t \nabla \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \boldsymbol{y}_i)$$

4

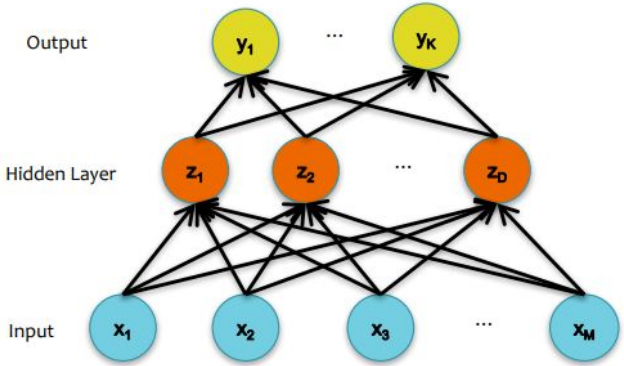
Slides taken from https://www.cs.cmu.edu/~mgormley/courses/10601-s17/slides/lecture21-cnn.pdf

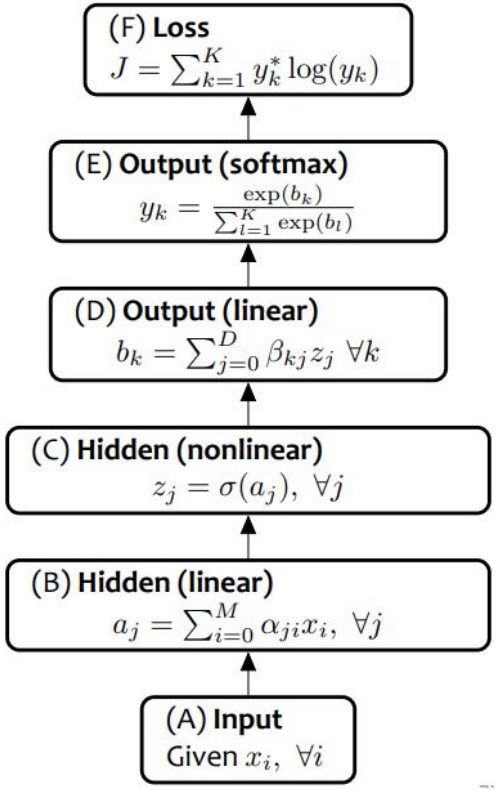
Convolutional neural network (CNN)

Multi-Class Output

Softmax Layer:

$$y_k = \frac{\exp(b_k)}{\sum_{l=1}^{K} \exp(b_l)}$$





Slides taken from https://www.cs.cmu.edu/~mgormley/courses/10601-s17/slides/lecture21-cnn.pdf

CNN: Fully Connected Layers

Fully-connected layers, also known as linear layers, connect every input neuron to every output neuron and are commonly used in neural networks. Fully connected layers enable the network to combine all the learned features from previous layers, especially after convolutional layers. This helps in abstracting and combining information to make a final decision or prediction

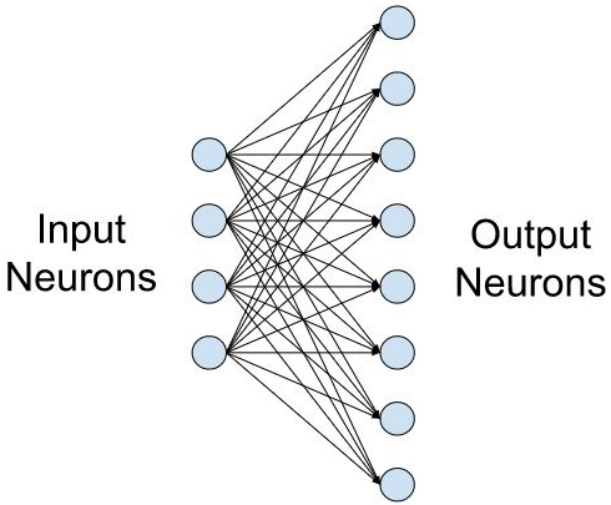


image source https://docs.nvidia.com/deeplearning/performance/dl-performance-fully-connected/index.htm

CNN: Sigmoid function

The Sigmoid activation function takes in any *real* number as the input and maps it to a number between 0 and 1. This is exactly why it's well-suited for binary classification.
The sigmoid function is defined as

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

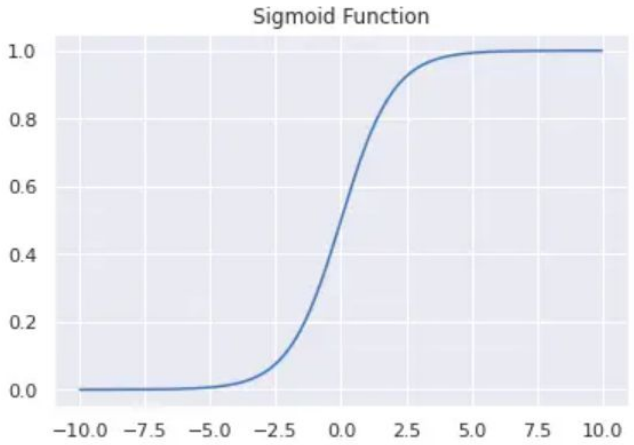


image source https://www.pinecone.io/learn/softmax-activation/

CNN: Softmax function

The softmax activation function takes in a vector of raw outputs of the neural network and returns a vector of probability scores. The softmax function is commonly used for multi-class classification. The softmax function is defined as

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$$

Let's consider a raw vector = [0.25, 1.23, -0.8]
Let's calculate the exponentials of each value

$$e^{0.25} \approx 1.284$$

$$e^{1.23} \approx 3.421$$

$$e^{-0.8} \approx 0.449$$

Sum of exponential values = 1.284 + 3.421 + 0.449 = 5.154

CNN: Softmax function cont.

The first element of the vector after softmax= 1.284 / 5.154 = 0.249
The second element of the vector after softmax = 3.421 / 5.154 = 0.664
The third element of the vector after softmax = 0.449/5.154 = 0.87

The vector after performing the softmax is [0.249, 0.664, 0.87]

CNN: ReLU function

ReLU is a powerful activation function that helps the network decide which parts of the input are important by "activating" some values and ignoring others.

$$\text{ReLU}(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

For example

- If the input is 3, ReLU outputs 3 (because it's positive).
- If the input is -2, ReLU outputs 0 (because it's negative).

ReLU adds non-linearity to the neural network.
ReLU is not differentiable at 0.
But we can get around that by defining a subgradient.

Remember we need activation functions to be differentiable for the weight update via gradient descent rule.

CNN: Loss functions

There are different Loss functions:

1. Binary Cross Entropy (BCE) Loss (for a two class problem)
2. Cross entropy (CE) loss (for multi-class problem)
3. Mean Square Error (MSE) Loss

We will try to look at the loss functions and backpropagation method in some other class.

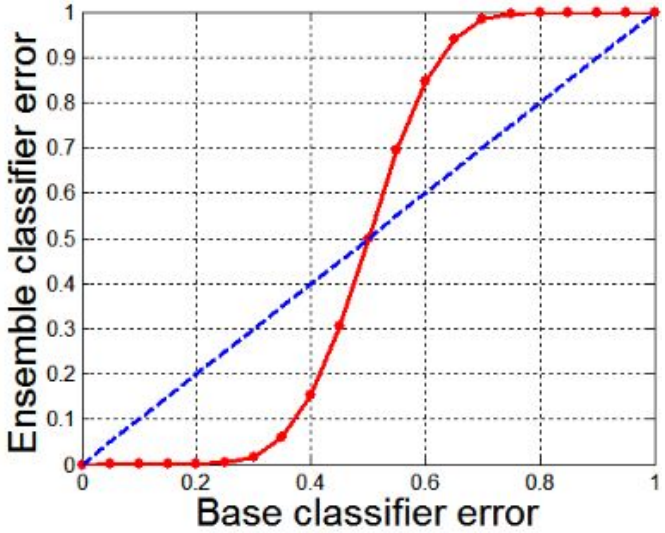
Ensemble Methods

Ensemble Methods

- Construct a set of classifiers from the training data
- Predict class label of test records by combining the predictions made by multiple classifiers

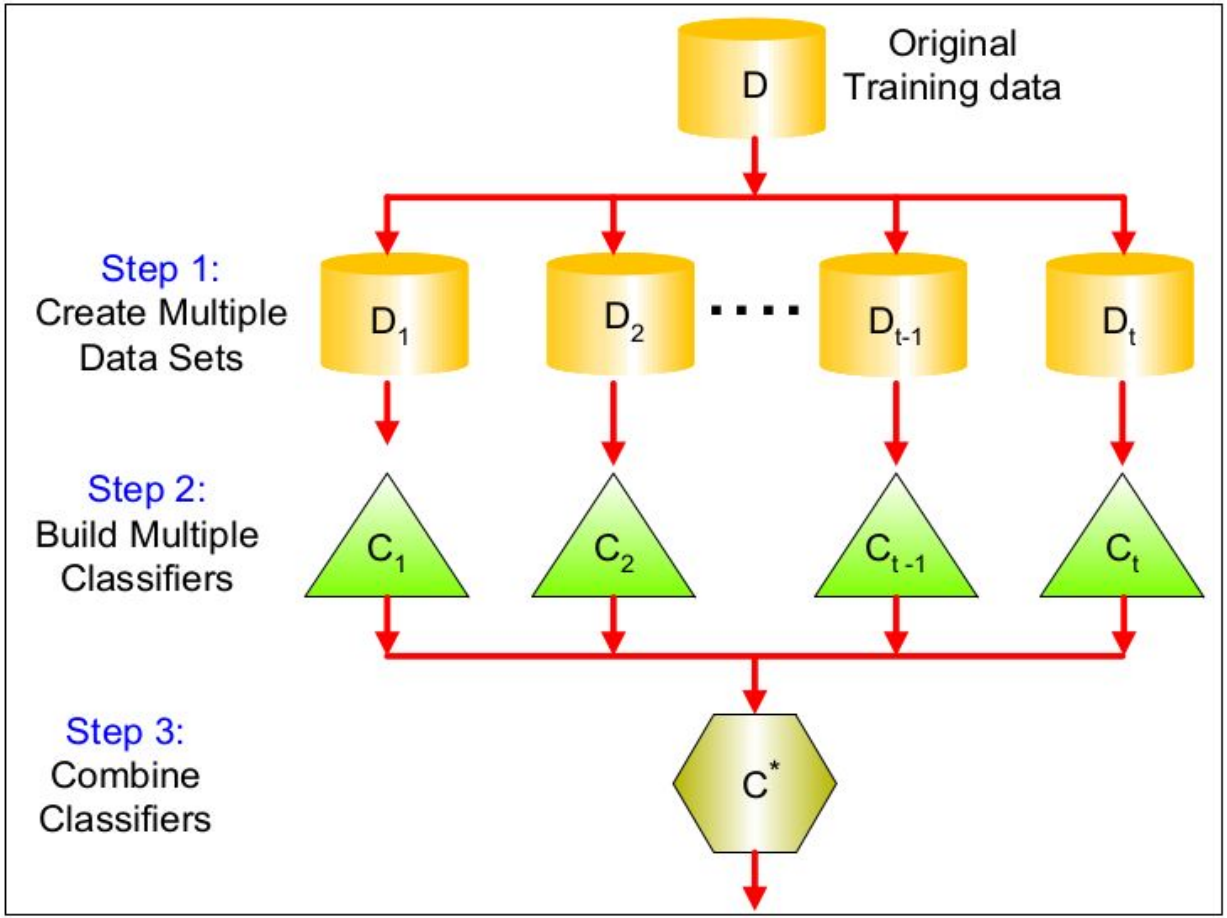
Why Ensemble Methods work?

- Suppose there are 25 base classifiers
  - Each classifier has error rate, $\varepsilon = 0.35$
  - Assume errors made by classifiers are uncorrelated
  - Probability that the ensemble classifier makes a wrong prediction:



$$P(X \geq 13) = \sum_{i=13}^{25} \binom{25}{i} \varepsilon^i (1-\varepsilon)^{25-i} = 0.06$$

General Approach



Types of Ensemble Method

- Manipulate data distribution
  - Example: bagging, boosting
- Manipulate input features
  - Example: random forests
- Manipulate class labels
  - Example: error-correcting output coding

Bagging

- Sampling with replacement

Original Data	1	2	3	4	5	6	7	8	9	10
Bagging (Round 1)	7	8	10	8	2	5	10	10	5	9
Bagging (Round 2)	1	4	9	1	2	3	2	7	3	2
Bagging (Round 3)	1	8	5	10	5	5	9	6	3	7

- Build classifier on each bootstrap sample

Bagging Algorithm

Algorithm 5.6 Bagging Algorithm

1: Let k be the number of bootstrap samples.
2: **for** $i = 1$ to k **do**
3: Create a bootstrap sample of size n, D_i.
4: Train a base classifier C_i on the bootstrap sample D_i.
5: **end for**
6: $C^*(x) = \arg\max_y \sum_i \delta\big(C_i(x) = y\big)$, $\{\delta(\cdot) = 1$ if its argument is true, and 0 otherwise.$\}$

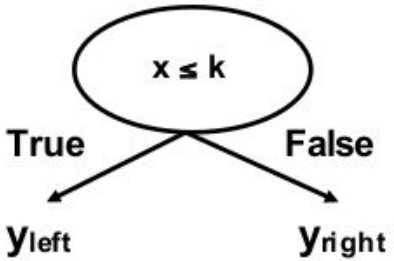
Bagging Example

- Consider 1-dimensional data set:

Original Data:

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
y	1	1	1	-1	-1	-1	-1	1	1	1

- Classifier is a decision stump
  - Decision rule: $x \leq k$ versus $x > k$
  - Split point k is chosen based on entropy



Bagging Example

Bagging Round 1:

x	0.1	0.2	0.2	0.3	0.4	0.4	0.5	0.6	0.9	0.9
y	1	1	1	1	-1	-1	-1	-1	1	1

x <= 0.35 ➔ y = 1
x > 0.35 ➔ y = -1

Bagging Example

Bagging Round 1:

x	0.1	0.2	0.2	0.3	0.4	0.4	0.5	0.6	0.9	0.9
y	1	1	1	1	-1	-1	-1	-1	1	1

x <= 0.35 ➔ y = 1
x > 0.35 ➔ y = -1

Bagging Round 2:

x	0.1	0.2	0.3	0.4	0.5	0.5	0.9	1	1	1
y	1	1	1	-1	-1	-1	1	1	1	1

x <= 0.7 ➔ y = 1
x > 0.7 ➔ y = 1

Bagging Round 3:

x	0.1	0.2	0.3	0.4	0.4	0.5	0.7	0.7	0.8	0.9
y	1	1	1	-1	-1	-1	-1	-1	1	1

x <= 0.35 ➔ y = 1
x > 0.35 ➔ y = -1

Bagging Round 4:

x	0.1	0.1	0.2	0.4	0.4	0.5	0.5	0.7	0.8	0.9
y	1	1	1	-1	-1	-1	-1	-1	1	1

x <= 0.3 ➔ y = 1
x > 0.3 ➔ y = -1

Bagging Round 5:

x	0.1	0.1	0.2	0.5	0.6	0.6	0.6	1	1	1
y	1	1	1	-1	-1	-1	-1	1	1	1

x <= 0.35 ➔ y = 1
x > 0.35 ➔ y = -1

Bagging Example

Bagging Round 6:

x	0.2	0.4	0.5	0.6	0.7	0.7	0.7	0.8	0.9	1
y	1	-1	-1	-1	-1	-1	-1	1	1	1

x <= 0.75 ➔ y = -1
x > 0.75 ➔ y = 1

Bagging Round 7:

x	0.1	0.4	0.4	0.6	0.7	0.8	0.9	0.9	0.9	1
y	1	-1	-1	-1	-1	1	1	1	1	1

x <= 0.75 ➔ y = -1
x > 0.75 ➔ y = 1

Bagging Round 8:

x	0.1	0.2	0.5	0.5	0.5	0.7	0.7	0.8	0.9	1
y	1	1	-1	-1	-1	-1	-1	1	1	1

x <= 0.75 ➔ y = -1
x > 0.75 ➔ y = 1

Bagging Round 9:

x	0.1	0.3	0.4	0.4	0.6	0.7	0.7	0.8	1	1
y	1	1	-1	-1	-1	-1	-1	1	1	1

x <= 0.75 ➔ y = -1
x > 0.75 ➔ y = 1

Bagging Round 10:

x	0.1	0.1	0.1	0.1	0.3	0.3	0.8	0.8	0.9	0.9
y	1	1	1	1	1	1	1	1	1	1

x <= 0.05 ➔ y = 1
x > 0.05 ➔ y = 1

Bagging Example

- Summary of Training sets:

Round	Split Point	Left Class	Right Class
1	0.35	1	-1
2	0.7	1	1
3	0.35	1	-1
4	0.3	1	-1
5	0.35	1	-1
6	0.75	-1	1
7	0.75	-1	1
8	0.75	-1	1
9	0.75	-1	1
10	0.05	1	1

Bagging Example

- Assume test set is the same as the original data
- Use majority vote to determine class of ensemble classifier

Round	x=0.1	x=0.2	x=0.3	x=0.4	x=0.5	x=0.6	x=0.7	x=0.8	x=0.9	x=1.0
1	1	1	1	-1	-1	-1	-1	-1	-1	-1
2	1	1	1	1	1	1	1	1	1	1
3	1	1	1	-1	-1	-1	-1	-1	-1	-1
4	1	1	1	-1	-1	-1	-1	-1	-1	-1
5	1	1	1	-1	-1	-1	-1	-1	-1	-1
6	-1	-1	-1	-1	-1	-1	-1	1	1	1
7	-1	-1	-1	-1	-1	-1	-1	1	1	1
8	-1	-1	-1	-1	-1	-1	-1	1	1	1
9	-1	-1	-1	-1	-1	-1	-1	1	1	1
10	1	1	1	1	1	1	1	1	1	1
Sum	2	2	2	-6	-6	-6	-6	2	2	2
Sign	1	1	1	-1	-1	-1	-1	1	1	1

Predicted Class

Task

1.Perform two rounds of bagging on the dataset given below. Use entropy as a criterion to create the decision stump.

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
y	1	-1	-1	-1	-1	-1	-1	1	1	1