

COMP 540      HW 5  
Peiguang Wang, Xinran Zhou  
Due: 3/5/2018

## 1: Deep neural networks

1. Why do deep neural networks typically outperform shallow networks?

**Solution.** By using deep neural network and adding more layers, we can approximate function using less parameters. The deep network encodes a set of prior beliefs about the structure of the function we want to learn. Thus, the deep neural networks reduce the amount of data we should use to get a satisfying result.

2. What is leaky RELU activation and why is it used?

**Solution.** Leaky relu is basically based on relu activation function and tries to fix the 'dying' problem of relu. When  $x < 0$ , the leaky relu has a small slope instead of being zero.

The reason why we use leaky relu is that it can give a small constant gradient when the input falls in the region  $x < 0$ . So it can fix the problem of "dead relu".

3. In one or more sentences, and using sketches as appropriate, contrast: AlexNet, VGGNet, GoogleNet and ResNet. What is the one defining characteristic of each network?

**Solution.** AlexNet: AlexNet uses RELU activate function instead of sigmoid function for the first time. And it also introduce a new dropout layer in the network.

VGGNet: VGGNet consists of either 16 or 19 convolutional layers and has very uniform architecture.

GoogleNet: This module is based on several very small convolutions in order to drastically reduce the number of parameters.

ResNet: ResNet introduces a so called "shortcut connection" that skips one or more layers, which allow the gradients can be backprop to the first layers. This allows us to train a much deeper network up to 152 layers.

## 2: Decision trees, entropy and information gain

1. Show that  $H(S) \leq 1$  and that  $H(S) = 1$  when  $p = 0.5$ .

**Solution.** Since

$$H(q) = -q \log(q) - (1 - q) \log(1 - q)$$

the second derivative of the  $-H(q)$  is non-negative, so the negative entropy is convex. The  $H(q)$  is concave. The maximum can be obtained at  $\frac{\partial H}{\partial q} = 0$

$$\frac{\partial H}{\partial q} = -\log(q) + \log(1 - q)$$

thus we got  $q = 0.5$ , which means that  $p = 0.5$  and  $H(S) = 1$ .

Therefore,  $H(S) \leq 1$  and that  $H(S) = 1$  when  $p = 0.5$ .

2. Calculate the reduction in cost using misclassification rate, entropy, and Gini index for models A and B. Which is the preferred split (model A or model B) according to these cost calculations?

**Solution. Misclassification rate:**

$$error_A = \frac{100 + 100}{400 + 400} = 0.25$$

$$error_B = \frac{200}{400 + 400} = 0.25$$

**Entropy** For both A and B:

$$H(D) = 1$$

For A:

$$H(D_1) = H(D_2) = -0.75 \log(0.75) - 0.25 \log(0.25) = 0.811$$

$$g(D, A) = H(D) - 0.5H(D_1) - 0.5H(D_2) = 0.189$$

For B:

$$H(D_1) = -\frac{1}{3} \log\left(\frac{1}{3}\right) - \frac{2}{3} \log\left(\frac{2}{3}\right) = 0.913$$

$$H(D_2) = 0$$

$$g(D, B) = H(D) - 0.75H(D_1) - 0.25H(D_2) = 0.312$$

**Gini Index:**

$$Gini(A) = 0.5(1 - 0.75^2 - 0.25^2) + 0.25(1 - 0.25^2 - 0.75^2) = 0.375$$

$$Gini(B) = 0.75(1 - \frac{2^2}{3} - \frac{1^2}{3}) + 0.25(1 - 1 - 0) = \frac{1}{3}$$

Among these three cost calculations, the entropy is the preferred split since the difference between A and B in this cost calculation is the biggest.

3. Can the misclassification rate ever increase when splitting on a feature? If so, give an example. If not, give a proof.

**Solution.** No, the misclassification rate will not increase when splitting on a feature.

### 3: Bagging

1. Assuming that the individual errors  $\epsilon_l(x)$  have zero mean and are uncorrelated, that is  $E_x[\epsilon_l(x)] = 0$  and  $E_x[\epsilon_m(x)\epsilon_l(x)] = 0$  for  $m \neq l$ , show that

$$E_{bag} = \frac{1}{L} E_{av}$$

**Solution.** Since

$$\epsilon_{bag} = \frac{1}{L} \sum_{l=1}^L (f(x) + \epsilon_l(x)) - f(x)$$

where  $\epsilon_l \sim N(\mu, \sigma_l^2)$ , and they are uncorrelated. If we calculate the  $E_{bag}$ , then

$$E_{bag} = E[\epsilon_{bag}(x)^2] = var(\epsilon_{bag}(x))$$

the result is  $\frac{1}{L^2} \sum_{l=1}^L \sigma_L^2$ . And we have

$$E_{av} = \frac{1}{L} \sum_{l=1}^L E_x[\epsilon_l(x)^2]$$