CSC2515: Homework #2

Due on October 3

Roger Grosse

Thomas Hollis

Problem 1

Prove that the entropy $H(X)$ is non-negative.

Solution

By definition, probability mass function $p(x)$ cannot be negative as it is the distribution of probabilities that a given random variable is exactly equal to particular values. Thus, we can state:

$$1 \geq p(x) \geq 0 \tag{1}$$

In addition, we know that:

$$H(X) = \sum_x p(x) \log_2 \frac{1}{p(x)} \tag{2}$$

Equivalently as in (2), we have:

$$H(X) = -\sum_x p(x) \log_2 p(x) \tag{3}$$

Since we assume finite sums, we know:

$$\log_2(x) \leq 0 \quad \forall \quad 1 \leq x < 0 \tag{4}$$

$$\sum_x x \geq 0 \quad \forall \quad x \geq 0 \tag{5}$$

Therefore, from (4) and (5) we can say:

$$\sum_x p(x) \log_2 p(x) \leq 0 \quad \forall \quad 1 \geq p(x) \geq 0 \tag{6}$$

This is because we can use l'Hopital's rule to check when $p(x) = 0$, as follows:

$$\lim_{p(x)\to 0} \sum_x p(x) \log_2 p(x) = \sum_x \lim_{p(x)\to 0} \frac{\frac{\mathrm{d}}{\mathrm{d}x}(\log_2 p(x))}{\frac{\mathrm{d}}{\mathrm{d}x}(\frac{1}{p(x)})} = 0 \tag{7}$$

Adding a minus sign in (6) we can state:

$$-\sum_x p(x) \log_2 p(x) \geq 0 \quad \forall \quad 1 \geq p(x) \geq 0 \tag{8}$$

Hence, combining (8), (3) and (1) we can state that:

$$H(X) \geq 0 \tag{9}$$

Therefore, we have proved that entropy is always non-negative.

Problem 2

Prove that $KL(p||q)$ is non-negative.

Solution

We know that:

$$KL(p||q) = \sum_x p(x) \log_2 \frac{q(x)}{p(x)} \tag{10}$$

Thus, let us define a function $f(x)$ as follows:

$$f(x) = -\log_2 \frac{q(x)}{p(x)} \tag{11}$$

Since this function $f(x)$ is concave on the set of positive real numbers (as stated in the Appendix), we can use the Jensen Inequality to state the following:

$$\mathbb{E}_p[f(x)] \geq f(\mathbb{E}_p[x]) \tag{12}$$

Since the expected value is the sum of the product of the function with its distribution, thus we can state:

$$\sum_x f(x)p(x) \geq -\log_2 \frac{q(\mathbb{E}_p[x])}{p(\mathbb{E}_p[x])} \tag{13}$$

Since the distribution of a constant is always 1, we can state:

$$\sum_x f(x)p(x) \geq -\log_2(1) \tag{14}$$

Substituting $f(x)$ from (11) into (14) yields:

$$\sum_x -\log_2 \frac{q(x)}{p(x)} p(x) \geq -\log_2(1) \tag{15}$$

Hence by taking the minus into the logarithm:

$$\sum_x \log_2 \frac{p(x)}{q(x)} p(x) \geq 0 \tag{16}$$

Thus, we have proved that $KL(p||q)$ is non-negative.

Problem 3

The Information Gain or Mutual Information between X and Y is $I(Y;X) = H(Y) - H(Y|X)$. Show that $I(Y;X) = KL(p(x,y)||p(x)p(y))$, where $p(x) = \Sigma_y p(x,y)$ is the marginal distribution of X.

Solution

We are trying to show that:

$$I(Y;X) = KL(p(x,y)||p(x)p(y)) \tag{17}$$

And we know that:

$$I(Y;X) = H(Y) - H(Y|X) \tag{18}$$

$$KL(p(x,y)||p(x)p(y)) = \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{19}$$

Thus:

$$I(Y;X) = -\sum_y p(y) \log_2 p(y) - \sum_x p(x) H(Y|X=x) \tag{20}$$

$$I(Y;X) = -\sum_y \log_2 p(y) \Big(\sum_x p(x,y) \Big) - \sum_x p(x) \Big(\sum_y p(y|x) \log_2 p(y|x) \Big) \tag{21}$$

$$I(Y;X) = -\sum_{x,y} p(x,y) \log_2 p(y) + \sum_{x,y} p(x) p(y|x) \log_2 p(y|x) \tag{22}$$

$$I(Y;X) = -\sum_{x,y} p(x,y) \log_2 p(y) + \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)} \tag{23}$$

$$I(Y;X) = \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{24}$$

Hence, substituting (24) into (19) we can state:

$$I(Y;X) = KL(p(x,y)||p(x)p(y)) \tag{25}$$

Thus we have proven the required statement.

Problem 3 continued on next page...

Problem 4

Consider m estimators $h_1, \cdots, h_m$, each of which accepts an input x and produces an output y, i.e., $y_i = h_i(x)$. These estimators might be generated through a Bagging procedure, but that is not necessary to the result that we want to prove. Consider the squared error loss function $L(y,t) = \frac{1}{2}(y-t)^2$. Show that the loss of the average estimator is smaller than the average loss of the estimators.

Solution

We know that the squared error loss function is:

$$L(y,t) = \frac{1}{2}(y-t)^2 \tag{26}$$

And we also know from Jensens inequality that for all convex functions:

$$\Phi(\mathbb{E}[X]) \leq \mathbb{E}[\Phi(X)] \tag{27}$$

Hence, since the squared error loss function is a convex function thus we can state by visual inspection that:

$$L(\bar{h}(x), t) \leq \frac{1}{m}\sum_{i=1}^{m} L(h_i(x), t) \tag{28}$$

The above is true since the loss function is a convex quadratic, $\bar{h}(x)$ is the expected value and the weighted sum is also a way of representing the expected value. This shows that the loss of the average estimator is smaller than the average loss of the estimators.

Problem 5

The goal of this question is to show that the AdaBoost algorithm changes the weights in order to force the weak learner to focus on difficult data points. Here we consider the case that the target labels are from the set {1, +1} and the weak learner also returns a classifier whose outputs belongs to {1,+1} (instead of {0,1}). Show that the error w.r.t. $(w', \cdots, w'_N)$ is exactly $\frac{1}{2}$. What is the interpretation of this result?

Solution

We know that:

$$err_t = \frac{\sum_{i=1}^{N} w_i \mathbb{I}\{h_t(\mathbf{x}^{(i)}) \neq t^{(i)}\}}{\sum_{i=1}^{N} w_i} \tag{29}$$

$$err'_t = \frac{\sum_{i=1}^{N} w'_i \mathbb{I}\{h_t(\mathbf{x}^{(i)}) \neq t^{(i)}\}}{\sum_{i=1}^{N} w'_i} \tag{30}$$

And we know that we can split the summations of (29) and (30) with two complimentary sets:

$$E = \{i : h_t(\mathbf{x}^{(i)}) \neq t^{(i)}\} \tag{31}$$

$$E^c = 1 - E = \{i : h_t(\mathbf{x}^{(i)}) = t^{(i)}\} \tag{32}$$

And since:

$$w'_i = w_i \exp\Big(- \alpha_t t^{(i)} h_t(\mathbf{x}^{(i)})\Big) \tag{33}$$

And since:

$$\alpha_t = \frac{1}{2} \log \frac{1 - err_t}{err_t} \tag{34}$$

Therefore we can say the updated weights for set E are:

$$w'_i = w_i \exp\Big(- \frac{1}{2} \log \frac{1 - err_t}{err_t} \times -1\Big) = w_i \sqrt{\frac{1 - err_t}{err_t}} \tag{35}$$

And we can say the updated weights for the complimentary set E^c are:

$$w'_i = w_i \exp\Big(- \frac{1}{2} \log \frac{1 - err_t}{err_t} \times 1\Big) = w_i \sqrt{\frac{err_t}{1 - err_t}} \tag{36}$$

Therefore combining (29-30) and (31-32):

$$err_t = \frac{\sum_{i \in E} w_i}{\sum_{i=1}^{N} w_i} = \frac{\sum_{i \in E} w_i}{\sum_{i \in E} w_i + \sum_{i \in E^c} w_i} \tag{37}$$

$$err'_t = \frac{\sum_{i \in E} w'_i}{\sum_{i \in E} w'_i + \sum_{i \in E^c} w'_i} \tag{38}$$

Substituting (35) and (36) into (38) yields:

$$err'_t = \frac{\sum_{i \in E} w_i \sqrt{\frac{1-err_t}{err_t}}}{\sum_{i \in E} w_i \sqrt{\frac{1-err_t}{err_t}} + \sum_{i \in E^c} w_i \sqrt{\frac{err_t}{1-err_t}}} \tag{39}$$

Problem 5 continued on next page...

Simplifying (39) yields:

$$err'_t = \frac{\sum_{i \in E} w_i}{\sum_{i \in E} w_i + \sum_{i \in E^c} w_i \frac{err_t}{1 - err_t}} \tag{40}$$

Substituting (37) into (40) yields:

$$err'_t = \frac{\sum_{i \in E} w_i}{\sum_{i \in E} w_i + \frac{\sum_{i \in E^c} w_i \frac{\sum_{i \in E} w_i}{\sum_{i \in E} w_i + \sum_{i \in E^c} w_i}}{1 - \frac{\sum_{i \in E} w_i}{\sum_{i \in E} w_i + \sum_{i \in E^c} w_i}}} \tag{41}$$

Simplifying (41) by multiplying top and bottom of the denominator fraction yields:

$$err'_t = \frac{\sum_{i \in E} w_i}{\sum_{i \in E} w_i + \frac{\sum_{i \in E^c} w_i \times \sum_{i \in E} w_i}{\sum_{i \in E} w_i + \sum_{i \in E^c} w_i - \sum_{i \in E} w_i}} = \frac{\sum_{i \in E} w_i}{\sum_{i \in E} w_i + \frac{\sum_{i \in E^c} w_i \times \sum_{i \in E} w_i}{\sum_{i \in E^c} w_i}} \tag{42}$$

Cancelling out in the denominator of (42) yields:

$$err'_t = \frac{\sum_{i \in E} w_i}{\sum_{i \in E} w_i + \sum_{i \in E} w_i} = \frac{1}{2} \tag{43}$$

Hence we have proven that the error with respect to the updated weights is exactly $\frac{1}{2}$.

The interpretation of this result is that AdaBoost indeed updates the weights in such a way as to force the weak learner to focus on difficult datapoints. Since a random algorithm would perform with a classification error of around one half, we expect the weak learner at iteration t to have an error of less than one half. However, after updating the weights at iteration t the weak learner now has an error of exactly one half. This means that the weights of the mislabelled data (difficult datapoints) were increased as to bump up the weak learner's error to exactly one half. In the next iteration at $t + 1$, the next weak learner is fed the incoming data with the updated weights (biased toward the difficult incorrectly labelled datapoints). His own error is expected to be less than one half until the weights are updated once again for the subsequent weak learner at iteration $t + 2$ and so on for all iterations up till $t + N$. This enables each subsequent learner to focus on the errors of the learners before him due to the weight update rule.

It is worth noting that the reason this occurs is due to the design of the classifier coefficient α. Indeed α is chosen such that it minimises the exponential error function of AdaBoost.