Homework 2

Deadline: Friday, Feb. 1, at 11:59pm.

Submission: You need to submit one file through MarkUs
`https://markus.teach.cs.toronto.edu/csc411-2019-01`

- Your answers to Questions 1, 2, and 3 as a PDF file titled `hw2_writeup.pdf`. You can produce the file however you like (e.g. LaTeX, Microsoft Word, scanner), as long as it is readable.

Neatness Point: One of the 10 points will be given for neatness. You will receive this point as long as we don't have a hard time reading your solutions or understanding the structure of your code.

Late Submission: 10% of the marks will be deducted for each day late, up to a maximum of 3 days. After that, no submissions will be accepted.

Weekly homeworks are individual work. See the Course Information handout `http://www.cs.toronto.edu/~mren/teach/csc411_19s/syllabus.pdf` for detailed policies.

1. [**4pts**] **Information Theory.** The goal of this question is to help you become more familiar with the basic equalities and inequalities of information theory. They appear in many contexts in machine learning and elsewhere, so having some experience with them is quite helpful. We review some concepts from information theory, and ask you a few questions.

   Recall the definition of the entropy of a discrete random variable X with probability mass function p: $H(X) = \sum_x p(x) \log_2 \left(\frac{1}{p(x)} \right)$. Here the summation is over all possible values of $x \in \mathcal{X}$, which (for simplicity) we assume is finite. For example, $\mathcal{X}$ might be $\{1, 2, \ldots, N\}$. You may assume $p(x) \log_2 (\left(\frac{1}{p(x)} \right) = 0$ if $p(x) = 0$.

   (a) [**1pt**] Prove that the entropy $H(X)$ is non-negative.

   An important concept in information theory is the relative entropy or the KL-divergence of two distributions p and q. It is defined as

   $$\mathsf{KL}(p||q) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)}.$$

   The KL-divergence is one of the most commonly used measure of difference (or divergence) between two distributions, and it regularly appears in information theory, machine learning, and statistics. For this question, you may assume $p(x) > 0$ and $q(x) > 0$ for all x.

   If two distributions are close to each other, their KL divergence is small. If they are exactly the same, their KL divergence is zero. KL divergence is not a true distance metric (since it isn't symmetric and doesn't satisfy the triangle inequality), but we often use it as a measure of dissimilarity between two probability distributions.

   (b) [**2pt**] Prove that $\mathsf{KL}(p||q)$ is non-negative. *Hint: you may want to use Jensen's Inequality, which is described in the Appendix.*

(c) **[1pt]** The Information Gain or Mutual Information between X and Y is $I(Y;X) = H(Y) - H(Y|X)$. Show that

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

where $p(x) = \sum_y p(x,y)$ is the marginal distribution of X.

2. **[2pts] Benefit of Averaging.** Consider m estimators $h_1, \dots, h_m$, 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

$$\bar{h}(x) = \frac{1}{m}\sum_{i=1}^{m} h_i(x),$$

is smaller than the average loss of the estimators. That is, for any x and t, we have

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

Hint: you may want to use Jensen's Inequality, which is described in the Appendix.

3. **[3pts] AdaBoost.** 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 $\{0,1\}$). Consider the t-th iteration of AdaBoost, where the weak learner is

$$h_t \leftarrow \underset{h \in \mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^{N} w_i \mathbb{I}\{h(\mathbf{x}^{(i)}) \neq t^{(i)}\},$$

the w-weighted classification error is

$$\text{err}_t = \frac{\sum_{i=1}^{N} w_i \mathbb{I}\{h_t(\mathbf{x}^{(i)}) \neq t^{(i)}\}}{\sum_{i=1}^{N} w_i},$$

and the classifier coefficient is $\alpha_t = \frac{1}{2}\log\frac{1-\text{err}_t}{\text{err}_t}$. (Here, log denotes the natural logarithm.) AdaBoost changes the weights of each sample depending on whether the weak learner h_t classifies it correctly or incorrectly. The updated weights for sample i is denoted by w_i' and is

$$w_i' \leftarrow w_i \exp\left(-\alpha_t t^{(i)} h_t(\mathbf{x}^{(i)})\right).$$

Show that the error w.r.t. $(w_1', \dots, w_N')$ is exactly $\frac{1}{2}$. That is, show that

$$\text{err}_t' = \frac{\sum_{i=1}^{N} w_i' \mathbb{I}\{h_t(\mathbf{x}^{(i)}) \neq t^{(i)}\}}{\sum_{i=1}^{N} w_i'} = \frac{1}{2}.$$

Note that here we use the weak learner of iteration t and evaluate it according to the new weights, which will be used to learn the $t+1$-st weak learner. What is the interpretation of this result?

Appendix: Convexity and Jensen's Inequality. Here, we give some background on convexity which you may find useful for some of the questions in this assignment. You may assume anything given here.

Convexity is an important concept in mathematics with many uses in machine learning. We briefly define convex set and function and some of their properties here. Using these properties are useful in solving some of the questions in the rest of this homework. If you are interested to know more about convexity, refer to Boyd and Vandenberghe, Convex Optimization, 2004.

A set C is *convex* if the line segment between any two points in C lies within C, i.e., if for any $x_1, x_2 \in C$ and for any $0 \le \lambda \le 1$, we have

$$\lambda x_1 + (1 - \lambda) x_2 \in C.$$

For example, a cube or sphere in $\mathbb{R}^d$ are convex sets, but a cross (a shape like X) is not.

A function $f : \mathbb{R}^d \to \mathbb{R}$ is *convex* if its domain is a convex set and if for all x_1, x_2 in its domain, and for any $0 \le \lambda \le 1$, we have

$$f(\lambda x_1 + (1 - \lambda) x_2) \le \lambda f(x_1) + (1 - \lambda) f(x_2).$$

This inequality means that the line segment between $(x_1, f(x_1))$ and $(x_2, f(x_2))$ lies above the graph of f. A convex function looks like $\smile$. We say that f is *concave* if $-f$ is convex. A concave function looks like $\frown$.

Some examples of convex and concave functions are (you do not need to use most of them in your homework, but knowing them is useful):

- Powers: x^p is convex on the set of positive real numbers when $p \ge 1$ or $p \le 0$. It is concave for $0 \le p \le 1$.
- Exponential: e^{ax} is convex on $\mathbb{R}$, for any $a \in \mathbb{R}$.
- Logarithm: $\log(x)$ is concave on the set of positive real numbers.
- Norms: Every norm on $\mathbb{R}^d$ is convex.
- Max function: $f(x) = \max\{x_1, x_2, \ldots, x_d\}$ is convex on $\mathbb{R}^d$.
- Log-sum-exp: The function $f(x) = \log(e^{x_1} + \ldots + e^{x_d})$ is convex on $\mathbb{R}^d$.

An important property of convex and concave functions, which you may need to use in your homework, is *Jensen's inequality.* Jensen's inequality states that if $\phi(x)$ is a convex function of x, we have

$$\phi(\mathbb{E}[X]) \le \mathbb{E}[\phi(X)].$$

In words, if we apply a convex function to the expectation of a random variable, it is less than or equal to the expected value of that convex function when its argument is the random variable. If the function is concave, the direction of the inequality is reversed.

Jensen's inequality has a physical interpretation: Consider a set $\mathcal{X} = \{x_1, \ldots, x_N\}$ of points on $\mathbb{R}$. Corresponding to each point, we have a probability $p(x_i)$. If we interpret the probability as mass, and we put an object with mass $p(x_i)$ at location $(x_i, \phi(x_i))$, then the centre of gravity of these objects, which is in $\mathbb{R}^2$, is located at the point $(\mathbb{E}[X], \mathbb{E}[\phi(X)])$. If ϕ is convex $\smile$, the centre of gravity lies above the curve $x \mapsto \phi(x)$, and vice versa for a concave function $\frown$.