Assignment 1: CS 215

Due: 14th August before 11:55 pm

Remember the honor code while submitting this (and every other) assignment. All members of the group should work on all parts of the assignment. We will adopt a zero-tolerance policy against any violation.

Submission instructions:

1. You should type out all the answers to the written problems in Word (with the equation editor) or using Latex, or write it neatly on paper and scan it. In either case, prepare a pdf file.
2. Put the pdf file and the code for the programming parts all in one zip file. The pdf should contain the names and ID numbers of all students in the group within the header. The pdf file should also contain instructions for running your code. Name the zip file as follows: A1-IdNumberOfFirstStudent-IdNumberOfSecondStudent.zip. (If you are doing the assignment alone, the name of the zip file is A1-IdNumber.zip).
3. Upload the file on moodle BEFORE 11:55 pm on the due date (i.e. 14th August). We will nevertheless allow and not penalize any submission until 2:00 am on the following day (i.e. 15th August). No assignments will be accepted thereafter.
4. Note that only one student per group should upload their work on moodle.
5. Please preserve a copy of all your work until the end of the semester.

Questions:

1. Given data-points $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, we proved in class that $-1 \leq r(x,y) \leq 1$. Now, you should prove that if for all i, $y_i = ax_i + b$ where a and b are constants and $a < 0$, then $r(x,y) = -1$. Is the converse of this statement true? Prove or disprove. To disprove, you can just produce a single counter-example. [5+5=10 points]
   Answer: See scan.
2. Referring to the question above, consider for all $i, 1 \leq i \leq n$, we have $z_i = ax_i + b$ and $w_i = cy_i + d$ where $a \neq 0, c \neq 0$. Here a, b, c, d are constants. Then prove that $r(z,w) = r(x,y)$. [5 points]
   Answer: See scan.
3. We proved the one-sided Chebyshev's inequality in class, i.e. $\frac{N(S_k)}{n} \leq \frac{1}{1+k^2}$ where S_k is the set of all numbers x_i, $1 \leq i \leq n$ such that $x_i - \bar{x} \geq k\sigma$ where $\bar{x}$ and σ are the mean and standard deviation of the samples, and $k > 0$. Now consider T_k is the set of all numbers x_i, $1 \leq i \leq n$ such that $x_i - \bar{x} \leq -k\sigma$. Then prove from first principles that $\frac{N(T_k)}{n} \leq \frac{1}{1+k^2}$. This is a variant of the one-sided Chebyshev's inequality. Recall that given a set X, $N(X)$ refers to the cardinality of the set, i.e. the number of numbers/elements in that set. Now, combining the one-sided Chebyshev's inequality and its variant, derive an upper bound for $\frac{N(U_k)}{n}$ where U_k is the set of numbers x_i, $1 \leq i \leq n$ such that $|x_i - \bar{x}| \geq k\sigma$. How does this upper bound compare with that predicted by the two-sided Chebyshev's inequality for different values of k? Justify your answer. [7+3+5=15 points]

Answer: Go through the exact same proof as on page 30 of the textbook but you define $y_i = \bar{x} - x_i$ instead of $y_i = x_i - \bar{x}$. Now, by definition of U_k, we know that U_k contains all and only those elements which are either in T_k or S_k (there is no element in common to both S_k and T_k). Hence $N(U_k) = N(T_k) + N(S_k)$ and hence $\frac{N(U_k)}{n} \leq \frac{2}{1+k^2}$. Using two-sided inequality, our bound would have been $\frac{1}{k^2}$. Comparing $\frac{2}{1+k^2}$ with $\frac{1}{k^2}$, we see that $\frac{2}{1+k^2} \leq \frac{1}{k^2}$ for $k \leq 1$, otherwise $\frac{2}{1+k^2} > \frac{1}{k^2}$. Hence for $k > 1$, the two-sided upper bound $\frac{N(U_k)}{n}$ is tighter. For $k \leq 1$, the upper bound predicted by either method is trivial.

4. Generate a sine wave in MATLAB of the form $y = 5\sin(2x + \pi/3)$ where x ranges from -10 to 10 in steps of 0.1. Now randomly select 40 values in the array y (using MATLAB function 'randperm') and corrupt them by adding random values from 100 to 200 using the MATLAB function 'rand'. This will generate a corrupted sine wave which we will denote as z. Now your job is to filter z using the following steps.

   - Create a new array w_{median} to store the filtered sine wave.
   - For a value at index i in y, consider a neighborhood $N(i)$ consisting of $y(i)$, 8 values to its right and 8 values to its left. For indices near the left or right end of the array, you may not have 8 neighbors in one of the directions. In such a case, the neighborhood will contain fewer values.
   - Set $w_{median}(i)$ to the median of all the values in $N(i)$. Repeat this for every i.

   This process is called as 'moving median filtering', and will produce a filtered signal in the end. Repeat the entire procedure described here using the arithmetic mean instead of the median. This is called as 'moving average filtering'. Plot the original (i.e. clean) sine wave y, the corrupted sine wave z and the filtered sine wave using mean and median on the same figure in different colors. Introduce a legend on the plot (find out how to do this in MATLAB). Include an image of the plot in your report. Now compute and print the mean squared error between each result and the original clean sine wave. Which method (median or arithmetic mean) produced better results? Why? Explain in your report. [6+3+3=12 points]
   Answer: See code part4.m. The median filter produces a mean squared error which is an order of magnitude lower and works better because median is more resistant to outliers than the mean. The median filter does produce errors at the crest and trough of the sine wave, however.

5. You will find a country-wise listing of life expectancy at the following weblink (on the right side of the web-page): `http://www.worldlifeexpectancy.com/world-rankings-total-deaths`. The relevant data from the weblink is extracted and stored in the text file LifeExpectancy.txt in the homework folder. You can load the data into the MATLAB workspace using the function 'dlmread'. Write a MATLAB program to compute the fraction of the number of countries whose life-expectancy lies in the interval $[\mu - k\sigma, \mu + k\sigma]$ where μ and σ are the mean and standard deviation of the life-expectancy of all the countries in the world, and where $k \in \{\pm 1, \pm 2, \pm 3 \pm 4, \pm 5\}$. How does this fraction compare with that predicted by (two-sided) Chebyshev's inequality? Explain in your report. Plot a graph with $|k|$ on the X axis, and the computed and Chebyshev-predicted fractions on the Y axis - all on the same plot. Include the plot in your report. [5+2+3=10 points]
   Answer: See code part5.m. The lower bounds produced by Chebyshev are loose (too low) as compared to the experimental values.

6. Imagine that you have computed the mean and standard deviation of a very large set of numbers. Now, you decide to add another number to this set. Write a MATLAB function to update the previously computed mean and another MATLAB function to update the previously computed standard deviation. Note that you are not allowed to simply recompute the mean and standard deviation by looping through all the data. As you might have already guessed, you will need to derive a formula for this. Include the formula and its derivation in your report. Your MATLAB functions should be of the form function newMean = UpdateMean (OldMean, NewDataValue, N) and function newStd = UpdateStd (OldMean, OldStd, NewMean, NewDataValue, N). [8 points]
   Answer: See the functions updateMean.m and updateStd.m. The formulae are derived and scanned.

   Scan

One-Sided Chebyshev Inequality

For $k>0$ $\quad \frac{N(k)}{n} \le \frac{1}{1+k^2}$

Proof: $y_i = x_i - \bar{x}$, $i = 1 \dots n$, for any $b>0$ we have

$$\sum_i (y_i+b)^2 \ge \sum_{i:\, y_i \ge ks} (y_i+b)^2 \quad \because (y_i+b)^2 \ge 0$$

$$\ge \sum_{i:\, y_i \ge ks} (ks+b)^2 \quad \because ks, b > 0.$$

$$= N(k)\,(ks+b)^2$$

$$\sum_i (y_i+b)^2 = \sum_i (y_i^2 + 2by_i + b^2)$$

$$= \sum_i y_i^2 + 2b\sum_i y_i + nb^2$$

$$= (n-1)s^2 + nb^2, \quad \because \sum y_i = \sum x_i - \bar{x} = 0$$

$$N(k) \le \frac{(n-1)s^2 + nb^2}{(ks+b)^2} \quad \forall\, b>0$$

$\therefore$ let $b = s/k$

$$\frac{N(k)}{n} \le \frac{s^2 + s^2/k^2}{(ks + s/k)^2}$$

$$\Rightarrow \frac{N(k)}{n} \le \frac{k^2/s^2\,(s^2 + s^2/k^2)}{k^2/s^2\,(ks + s/k)^2}$$

$$\Rightarrow \boxed{\frac{N(k)}{n} \le \frac{k^2+1}{(k^2+1)^2} = \frac{1}{k^2+1}}$$

Q2

Page No :
Date : X \

Proof that $r(\{x_i\}, \{y_i\}) = r(\{ax_i+b\}, \{cy_i+d\})$.

Note that

$$r(\{ax_i+b\}, \{cy_i+d\})$$

$$= \frac{\sum_{i=1}^{N} \left((ax_i+b) - (a\bar{x}+b)\right)\left((cy_i+d) - (c\bar{y}+d)\right)}{\sqrt{\sum_{i=1}^{N} (ax_i+b-(a\bar{x}+b))^2 \sum_{i=1}^{N} (cy_i+d-(c\bar{y}+d))^2}}$$

$$= \frac{\sum_{i=1}^{N} a(x_i-\bar{x})\, c(y_i-\bar{y})}{ac\sqrt{\sum_{i=1}^{N} (x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{N} (y_i-\bar{y})^2}}$$

$$= \frac{\sum_{i=1}^{N} a(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i-\bar{x})^2 \sum_{i=1}^{N} (y_i-\bar{y})^2}} \quad (a \neq 0,\ c \neq 0)$$

If $a = 0$ or $c = 0$, then correlation coefficient is undefined.

Q1

Page No.:
Date:

Prod that if $y_i = a + bx_i$, then

$r(\{x_i\}, \{y_i\}) = +1$ if $b > 0$

$= -1$ if $b < 0$

$=$ undefined if $b = 0$.

$$r(\{x_i\}, \{y_i\})$$

$$= \frac{\sum_{i=1}^{N} (x_i - \bar{x})\left(bx_i + a - (b\bar{x} + a)\right)}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2 \sum_{i=1}^{N} (bx_i + a - b\bar{x} + a)^2}}$$

$$= \frac{\sum_{i=1}^{N} b(x_i - \bar{x})^2}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2} \sqrt{b^2 \sum_{i=1}^{N} (x_i - \bar{x})^2}}$$

$$= \frac{b}{\sqrt{b^2}} = -1 \text{ if } b < 0 \text{ as we take positive square root}$$

$$= +1 \text{ if } b > 0$$

$$= \text{undefined } (0/0)$$

Page No:
Date: \ \

Non-calculus proof that median minimizes total absolute deviation ie

$$\arg\min_y \sum_{i=1}^{n} |y - x_i| = \text{median}\left(\{x_i\}_{i=1}^{n}\right)$$

Proof: Considers two numbers x_1 and x_2. Let α be a number such that $x_1 \le \alpha \le x_2$.

Then $\sum_{i=1}^{2} |\alpha - x_i| = x_2 - \alpha + \alpha - x_1$

$= x_2 - x_1$.

If $\alpha < x_1$, then $\sum_{i=1}^{2} |\alpha - x_i| = x_1 + x_2 - 2\alpha$

$> x_1 + x_2 - 2x_1 = x_2 - x_1$

Also if $\alpha > x_2$, then $\sum_{i=1}^{2} |\alpha - x_i| = -(x_2 + x_1) + 2\alpha$

$> 2x_2 - (x_2 + x_1)$

$= x_2 - x_1$

Now consider numbers $x_1 \le x_2 \le \dots \le x_n$ and nested intervals $[x_1, x_n]$, $[x_2, x_{n-1}]$, $\dots$, $[x_c, x_{n+1-c}]$ where $c = n/2$ (n is even) or $(n+1)/2$ (n is odd). If n is odd, then the innermost interval is $[x_c, x_c]$ and x_c is the median.

For any interval $[x_i, x_j]$ we have seen that any number $\alpha \in [x_i, x_j]$ minimizes the total absolute deviation. So if we choose $\alpha \in \bigcap_{i=1}^{c} [x_i, x_{n+1-i}]$, then it will minimize $\sum_{i=1}^{c} |x_i - \alpha| + |x_{n+1-i} - \alpha|$

$= \sum_{i=1}^{n} |x_i - \alpha|$

When n is even, the innermost interval is

$[x_{n/2}, x_{n/2+1}]$ and hence any α s.t. $x_{n/2} \leq \alpha \leq x_{n/2+1}$ minimizes the total absolute deviation. And this α will lie inside each of the 'c' intervals and minimize the absolute deviation for those intervals - and hence the total absolute deviation.

The least value of the total absolute deviation is

$$\sum_{j=n+1-c}^{n} x_j - \sum_{j=1}^{c} x_j$$

Q6

Let $\bar{x}_j$ = arithmetic mean of first j numbers

$\bar{x}_{j+1}$ = " " " " $j+1$ "

$$\bar{x}_{j+1} = \frac{1}{j+1}\sum_{i=1}^{j+1} x_i = \frac{1}{j+1}\sum_{i=1}^{j}(x_i + x_{j+1})$$

$$= \frac{j}{j+1}\bar{x}_j + \frac{x_{j+1}}{j+1} = \bar{x}_j + \frac{x_{j+1} - \bar{x}_j}{j+1}$$

Thus to compute $\bar{x}_{j+1}$ you need to know only $\bar{x}_j$ and of course the new data value x_{j+1}. You do not need to loop over the entire dataset to find $\bar{x}_{j+1}$.

For the standard deviation, we have

$$j s_{j+1}^2 = \sum_{i=1}^{j}(x_i - \bar{x}_{j+1})^2 + (x_{j+1} - \bar{x}_{j+1})^2$$

$$= \sum_{i=1}^{j}\left(x_i - \bar{x}_j - \frac{x_{j+1} - \bar{x}_j}{j+1}\right)^2 + (x_{j+1} - \bar{x}_{j+1})^2$$

$$= \sum_{i=1}^{j}\left\{(x_i - \bar{x}_j)^2 + \left(\frac{x_{j+1} - \bar{x}_j}{j+1}\right)^2 - 2(x_i - \bar{x}_j)\frac{(x_{j+1} - \bar{x}_j)}{j+1}\right\}$$

$$+ (x_{j+1} - \bar{x}_{j+1})^2$$

where $\sum_{i=1}^{j}(x_i - \bar{x}_j)^2 = (j-1)s_j^2$

$$\therefore j\, s_{j+1}^2 = (j-1)\, s_j^2 + j\left(\frac{x_{j+1} - \overline{x_j}}{j+1}\right)^2$$

$$-2j\left(\sum_{i=1}^{j}(x_i - \overline{x_j})\right)\left(\frac{x_{j+1} - \overline{x_j}}{j+1}\right)$$

$\sum_{i=1}^{j}(x_i - \overline{x_j}) \longrightarrow 0$

$$+\left(x_{j+1} - \overline{x_{j+1}}\right)^2$$

$$= (j-1)\, s_j^2 + j\left(\frac{x_{j+1} - \overline{x_j}}{j+1}\right)^2 + \left(x_{j+1} - \overline{x_{j+1}}\right)^2$$

$$= (j-1)\, s_j^2 + j\left(\overline{x_{j+1}} - \overline{x_j}\right)^2 + \left(x_{j+1} - \overline{x_{j+1}}\right)^2$$

$\frac{x_{j+1} - \overline{x_j}}{j+1} \updownarrow \overline{x_{j+1}} - \overline{x_j}$

→ see formula for mean

$$\therefore s_{j+1}^2 = \left(1 - \frac{1}{j}\right) s_j^2 + \left(\overline{x_{j+1}} - \overline{x_j}\right)^2 + \frac{\left(x_{j+1} - \overline{x_{j+1}}\right)^2}{j}$$

This is the formula to update the standard deviation.