1 Linear Regression

1. T/F: Let r be the correlation coefficient between $\vec{x} = (x_1, \ldots, x_n)$ and $\vec{y} = (y_1, \ldots, y_n)$. Then $\arccos r$ is the angle between the vectors $\vec{x} - \bar{x}$ and $\vec{y} - \bar{y}$.

   True. Let $\vec{v} = \vec{x} - \bar{x}$ and $\vec{w} = \vec{y} - \bar{y}$. Then

   $$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} = \frac{\vec{v} \cdot \vec{w}}{|\vec{v}||\vec{w}|} = \cos\alpha,$$

   where α is the angle between $\vec{v}$ and $\vec{w}$. So $\arccos r = \alpha$.

2. Recall that the CBS inequality gives us

   $$|\vec{v} \cdot \vec{w}| \leq |\vec{v}| \cdot |\vec{w}|. \tag{1}$$

   (a) What does (1) tell us about the angle between $\vec{v}$ and $\vec{w}$?

   $$|\cos\alpha| = \frac{|\vec{v} \cdot \vec{w}|}{|\vec{v}||\vec{w}|} \leq 1.$$

   This holds for all α, so the inequality doesn't tell us anything about the angle between the two vectors. And it should not, since $\vec{v}$ and $\vec{w}$ are arbitrary vectors.

   (b) Let $\vec{x} = (x_1, \ldots, x_n)$, and $\vec{y} = (y_1, \ldots, y_n)$.

   Let $\vec{v} = (x_1 - \bar{x}, \ldots, x_n - \bar{x})$, and $\vec{w} = (y_1 - \bar{y}, \ldots, y_n - \bar{y})$.

   What does (1) tell us about the correlation coefficient r between $\vec{x}$ and $\vec{y}$?

   (1) gives us that

   $$\left|\sum(x_i - \bar{x})(y_i - \bar{y})\right| \leq \sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2},$$

   which means that

   $$|r| = \left|\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}\right| \leq 1.$$

   So (1) tells us that $|r| \leq 1$.

   (c) For what kind of vectors $\vec{v}$ and $\vec{w}$ is equality attained, when $\vec{v}$ and $\vec{w}$ are vectors with two components?

   Let $\vec{v} = (a_1, a_2)$, and $\vec{w} = (b_1, b_2)$. Squaring both sides of (1), we get

   $$(a_1b_1 + a_2b_2)^2 = (a_1^2 + a_2^2)(b_1^2 + b_2^2).$$

   Then

   $$(a_1b_1)^2 + (a_2b_2)^2 = (a_1b_1)^2 + (a_2b_1)^2 + (a_1b_2)^2 + (a_2b_2)^2.$$

   So $(a_2b_1)^2 + (a_1b_2)^2 = 0$. Since both terms are nonnegative, we have

   $$(a_2b_1)^2 = (a_1b_2)^2 (= 0).$$

If $b_2, a_2 \neq 0$, then we have

$$\frac{a_1}{a_2} = \frac{b_1}{b_2},$$

which means that $\vec{w}$ and $\vec{v}$ are multiples of one another.

If one of $\vec{v}$ or $\vec{w}$ is the zero vector, the equality also holds.

If neither is the zero vector, but are both horizontal or vertical vectors, the equality also holds. In this case $\vec{w}$ and $\vec{v}$ are also multiples of one another.

2 Needleman-Wunsch Algorithm

Move	$\rightarrow$ X _	$\downarrow$ _ X	$\searrow$ X Y	$\searrow$ X X
Score	0	0	0	1

1. T/F:
   (a) A diagonal move is the only move that could increase the total score.

   True.

   (b) Any diagonal move can increase the total score.

   False. Only diagonal moves that lead to two aligned nucleotides can increase the total score.

2. Translate the following alignment into a path. What's its score?

G	C	T	A	_
G	_	T	A	G

Path:

	G	C	T	A
G	$\searrow$	$\rightarrow$		
T			$\searrow$	
A				$\searrow$
G				$\downarrow$

Score: $1 + 1 + 1 = 3$

3. Translate the following path into an alignment. What's its score?

   Score: $1 + 1 = 2$.

4. (HW37 # 2 partial) Consider the following new scoring system in Table 1. Compare the two alignments. Which one is better?

	G	T	C	G
T	$\rightarrow$	$\searrow$		
A			$\searrow$	
G				$\searrow$
C				$\downarrow$

Alignment:

G T C G _

_ T A G C

new gap	gap extension	match	mismatch
-5	-1	1	-1

Table 1: New scoring system

G A A A A A A T G A A A A A A T

G _ _ A _ A _ T G A A _ _ _ _ T

The first alignment has a score of $4 + (-5) * 3 - 1 = -12$. The second alignment has a score of $4 + (-5) + (-3) = -4$. So the second one is better. Note that using our old scoring system, the two would receive the same score.

5. (HW37 # 3 partial)

   (a) Besides being square, what type of matrix is the similarity matrix in table 2?

   It is also a symmetric matrix, because $A^T = A$.

   (b) Write the similarity matrix for the alphabet below, and then compute its eigenvalues and eigenvectors.

   i. Alphabet "AC".

	A	C
A	1	-1
C	-1	1

$(1-\lambda)^2 - (-1)^2 = 0 \Rightarrow \lambda_1 = 0, \lambda_2 = 2.$

   ii. (Challenge) Alphabet "ACG".

	A	C	G
A	1	-1	-1
C	-1	1	-1
G	-1	-1	1

We have $(\lambda + 1)(\lambda - 2)^2 = 0$. So $\lambda_1 = -1$, and $\lambda_{2,3} = 2$.

	A	C	G	T
A	1	-1	-1	-1
C	-1	1	-1	-1
G	-1	-1	1	-1
T	-1	-1	-1	1

Table 2: Similarity matrix