Name: Hye Soo Choi **Student ID: 23274190**

CS 189: Introduction to Machine Learning

Homework 3

Due: March 3, 2016 at 11:59pm

Problem 1: Independence vs. Correlation.

(a) The joint probability density table of (X, Y) is drawn as below.

$X\backslash Y$	-1	0	1
-1	0	1/4	0
0	1/4	0	1/4
1	0	1/4	0

Therefore,

$$E[XY] = 0,$$

$$E[X] = 0, E[Y] = 0.$$

Since $E[XY] = E[X]E[Y]$, X and Y are uncorrelated. X and Y are not independent because

$$0 = P\{X = 0, Y = 0\} \neq P\{X = 0\}P\{Y = 0\} = \frac{1}{2} \cdot \frac{1}{2}.$$

(b) X, Y, Z are pairwise independent. This is because

$$P\{X = 0\} = P\{X = 1\} = P\{Y = 0\} = P\{Y = 1\} = P\{Z = 0\} = P\{Z = 1\} = \frac{1}{2},$$

and, no matter what value X might have, $Y|X$ always takes a value $\{0, 1\}$ with equal probability since B_3, which is independent of X, takes a value of $\{0, 1\}$ with equal probability. Therefore the conditional distribution of Y given X is the same as the original distribution of Y. Therefore X and Y are independent. By symmetry, we can easily prove that Y and Z, Z and X are pairwise independent as well.
Since it is always true that

$$X \oplus Y \oplus Z = (B_1 \oplus B_2) \oplus (B_2 \oplus B_3) \oplus (B_3 \oplus B_1) = (B_1 \oplus B_1) \oplus (B_2 \oplus B_2) \oplus (B_3 \oplus B_3) = 0 \oplus 0 \oplus 0 = 0,$$

X,Y,Z are not mutually independent. To be more specific,

$$0 = P\{X = 0, Y = 0, Z = 1\} \neq P\{X = 0\}P\{Y = 0\}P\{Z = 1\} = \frac{1}{8}.$$

Problem 4: Covariance Matrixes and Decompositions.

(a) The inverse of Σ_X will not exist if and only if(TFAE)

- Σ_X has determinant zero,
- Σ_X has at least one eigenvalue of zero,
- there exists nonzero $y \in R^N$ such that $y^\top \Sigma_X y = 0$,
- there exists nonzero $y \in R^N$ such that $E[(y^\top (X-\mu))^2] = 0$,
- there exists nonzero $y \in R^N$ such that $y^\top (X-\mu) = 0$ almost surely,
- there exists nonzero $y \in R^N$ such that $y^\top X$ is some constant almost surely,
- there exists some random variable X_i which can be expressed as a linear combination of other X_j's.

We can remove all the X_i's which are expressed as a linear combination of other X_j's and preserve only the smallest number of X_j's that span all X_i's. By doing so, we can transform X into X' whose $\Sigma_{X'}$ is invertible, without losing any information: By a linear combination, we are able to restore the removed elements X_i's always.

(b) Let's denote the spectral decomposition of Σ^{-1} as $UDU^\top$, where $D = \mathrm{diag}(\lambda_i)$ is a diagonal matrix along with the eigenvalues of Σ^{-1} and U is a matrix whose columns are corresponding normalized eigenvectors of length 1. Write $D^{\frac{1}{2}}$ as $\mathrm{diag}(\lambda_i^{\frac{1}{2}})$, then

$$x^\top \Sigma^{-1} x = x^\top U D^{\frac{1}{2}} D^{\frac{1}{2}} U^\top x = ||D^{\frac{1}{2}} U^\top x||_2^2.$$

It follows that $A = D^{\frac{1}{2}} U$.

(c) When we transform it to $||Ax||_2^2$, $x^\top \Sigma^{-1} x$ have intuitive meaning of a squared distance from origin after rotating x around origin, with the rotation matrix $U^\top$ and either stretching or contracting the rotated vector by size of eigenvalues. note that the rotation transforms all eigenvector onto a standard axis. By multiplying a diagonal matrix $D^{\frac{1}{2}}$, a vector is stretched or contracted along standard axis .

(d) Observe that

$$\min_{x:||x||_2=1} ||Ax||_2 = \min_{x:||x||_2=1} ||D^{\frac{1}{2}} U^\top x||_2 = \min_{x:||x||_2=1} ||D^{\frac{1}{2}} x||_2$$

$$\max_{x:||x||_2=1} ||Ax||_2 = \max_{x:||x||_2=1} ||D^{\frac{1}{2}} U^\top x||_2 = \max_{x:||x||_2=1} ||D^{\frac{1}{2}} x||_2.$$

Since $D^{\frac{1}{2}}$ is a diagonal matrix, the minimum of $||Ax||_2^2$ is just the square of the minimum diagonal values of $D^{\frac{1}{2}}$, that is, the minimum eigenvalue of Σ^{-1}. Similarly, the maximum of $||Ax||_2^2$ is just the square of the maximum diagonal values of $D^{\frac{1}{2}}$, that is, the maximum eigenvalue of Σ^{-1}. To maximize $f(x)$, we should minimize $x^\top \Sigma x$ and therefore we should choose the eigenvector that matches with the smallest eigenvalues of Σ^{-1}. This is because, to minimize $x^\top \Sigma x$, it should be that $U^\top x = e_i$, where λ_i is the smallest eigenvalue. Equivalently, $x = Ue_i$, the eigenvector that matches with λ_i.

If X_i's are pairwise independent, then covariance matrix Σ becomes a diagonal matrix whose diagonal elements are the variance of X_i's, and U is an N dimensional identity matrix. This implies that an eigenvalue λ_i of Σ^{-1} is equal to a inverse of $Var(X_i)$, for $1 \leq i \leq N$. Thus, the minimum of $||Ax||_2^2$ is the minimum of $\frac{1}{Var(X_i)}, 1 \leq i \leq N$, and likewise the maximum of $||Ax||_2^2$ is the maximum of $\frac{1}{Var(X_i)}, 1 \leq i \leq N$. To maximize $f(x)$, we should choose an elementary vector $e_{i'}$, where $i' = \text{argmax}_i Var(X_i)$.