HOMEWORK 8

>>Sean Sun<<
>>9078202463<<

Instructions: Although this is a programming homework, you only need to hand in a pdf answer file. There is no need to submit the latex source or any code. You can choose any programming language, as long as you implement the algorithm from scratch.
Use this latex file as a template to develop your homework. Submit your homework on time as a single pdf file to Canvas. Please check Piazza for updates about the homework.

1 Directed Graphical Model [20 points]

Consider the directed graphical model (aka Bayesian network) in Figure 1.

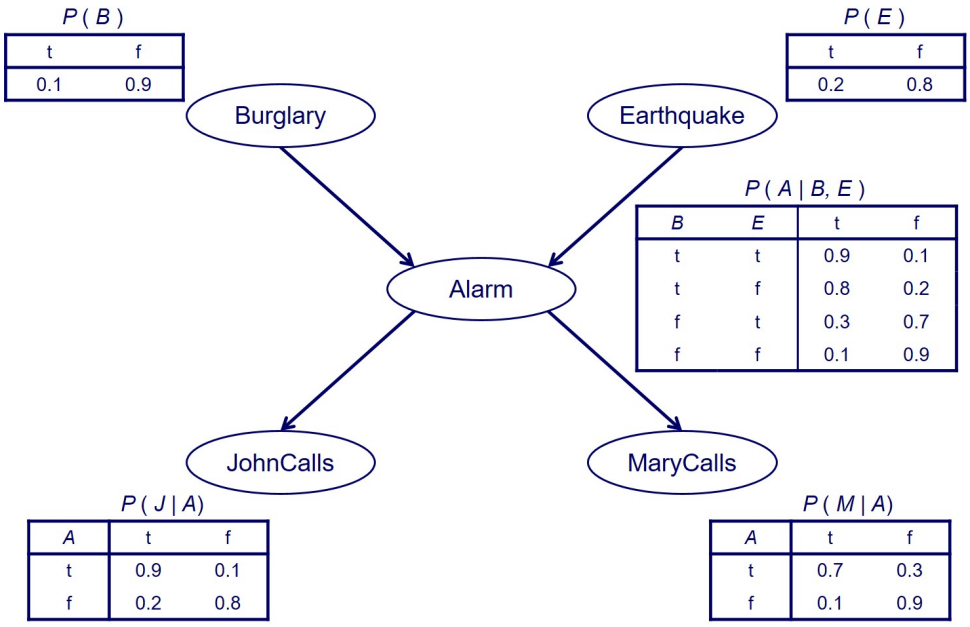


Figure 1: A Bayesian Network example.

Compute $P(B = t \mid E = f, J = t, M = t)$ and $P(B = t \mid E = t, J = t, M = t)$. These are the conditional probabilities of a burglar in your house (yikes!) when both of your neighbors John and Mary call you and say they hear an alarm in your house, but without or with an earthquake also going on in that area (what a busy day), respectively.

$$P(B = t \mid E = f, J = t, M = t)$$
$$= \frac{P(B=t,E=f,J=t,M=t)}{P(E=f,J=t,M=t)} = \frac{\sum_A P(B=t)P(E=f)P(M=t|A)P(J=t|A)P(A|B=t,E=f)}{\sum_A \sum_B P(B)P(E=f)P(M=t|A)P(J=t|A)P(A|B,E=f)} = \frac{0.04064}{0.09896}$$
$$= 0.4107$$

$$P(B = t \mid E = t, J = t, M = t)$$
$$= \frac{P(B=t,E=t,J=t,M=t)}{P(E=t,J=t,M=t)} = \frac{\sum_A P(B=t)P(E=t)P(M=t|A)P(J=t|A)P(A|B=t,E=t)}{\sum_A \sum_B P(B)P(E=t)P(M=t|A)P(J=t|A)P(A|B,E=t)} = \frac{0.01138}{0.04792}$$
$$= 0.2375$$

2 Chow-Liu Algorithm [20 pts]

Suppose we wish to construct a directed graphical model for 3 features X, Y, and Z using the Chow-Liu algorithm. We are given data from 100 independent experiments where each feature is binary and takes value T or F. Below is a table summarizing the observations of the experiment:

X	Y	Z	Count
T	T	T	36
T	T	F	4
T	F	T	2
T	F	F	8
F	T	T	9
F	T	F	1
F	F	T	8
F	F	F	32

1. Compute the mutual information $I(X, Y)$ based on the frequencies observed in the data.

   $I(X,Y) = \sum_{(x,y)} \hat{P}(X=x, Y=y) log \frac{\hat{P}(X=x,Y=y)}{\hat{P}(X=x)\hat{P}(Y=y)} = 0.2781$

2. Compute the mutual information $I(X, Z)$ based on the frequencies observed in the data.

   $I(X,Z) = \sum_{(x,z)} \hat{P}(X=x, Z=z) log \frac{\hat{P}(X=x,Z=z)}{\hat{P}(X=x)\hat{P}(Z=z)} = 0.1328$

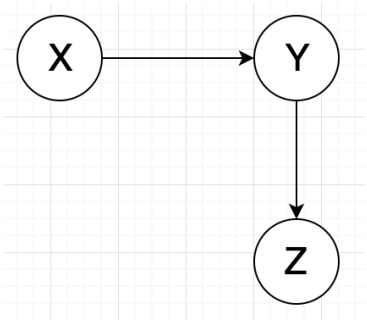
3. Compute the mutual information $I(Z, Y)$ based on the frequencies observed in the data.

   $I(Z,Y) = \sum_{(z,y)} \hat{P}(Z=z, Y=y) log \frac{\hat{P}(Z=z,Y=y)}{\hat{P}(Z=z)\hat{P}(Y=y)} = 0.3973$

4. Which undirected edges will be selected by the Chow-Liu algorithm as the maximum spanning tree?

   We should choose E(Z,Y) and E(X,Y) according to greedy algotirhm

5. Root your tree at node X, assign directions to the selected edges.



3 Kernel SVM [20 points]

Consider the following kernel function defined over $z, z' \in Z$:

$$k(z, z') = \begin{cases} 1 & \text{if } z = z', \\ 0 & \text{otherwise.} \end{cases}$$

1. Prove that for any integer $m > 0$, any $z_1, \ldots, z_m \in Z$, the $m \times m$ kernel matrix $K = [K_{ij}]$ is positive semi-definite, where $K_{ij} = k(z_i, z_j)$ for $i, j = 1 \ldots m$. Hint: An $m \times m$ matrix K is positive semi-definite if $\forall v \in \mathbb{R}^m, v^\top K v \geq 0$.

   K is an identity matrix, i.e.$K = I_m$
   So, $\forall v \in \mathbb{R}^m, v^\top K v = v^\top I_m v = v^\top v \geq 0$
   So, kernel matrix $K = [K_{ij}]$ is positive semi-definite

2. Given a training set $(z_1, y_1), \ldots, (z_n, y_n)$ with binary labels, the dual SVM problem with the above kernel k will have parameters $\alpha_1, \ldots, \alpha_n, b \in \mathbb{R}$. The predictor for input z takes the form

$$f(z) = \sum_{i=1}^{n} \alpha_i y_i k(z_i, z) + b.$$

Recall the label prediction is $\text{sgn}(f(z))$. Prove that there exists $\alpha_1, \ldots, \alpha_n, b$ such that f correctly separates the training set. In other words, k induces a feature space rich enough such that in it any training set can be linearly separated.
When a same z can have different y values, the bayes error will beb greater than zero, which means it is by no means we can have a set of parameters to linearly separate the training set. So, for this question, I assume all z values are distinct or same "z" values will have same y value.

For every z_j in training set, $\sum_{i=1}^{n} \alpha_i y_i k(z_i, z) = \alpha_j y_j k(z_j, z_j) = \alpha_j y_j$ since all other $k(z_i, z_j)$ are zeros.
So, $f(z_j) = \alpha_j y_j + b$
When we set all α to be 1 and set b to be 0, $f(z_j) = y_j$, which means we can linearly separate all training set points.
So, there exists $\alpha_1, \ldots, \alpha_n, b$ such that f correctly separates the training set.

3. How does that f predict input z that is not in the training set?

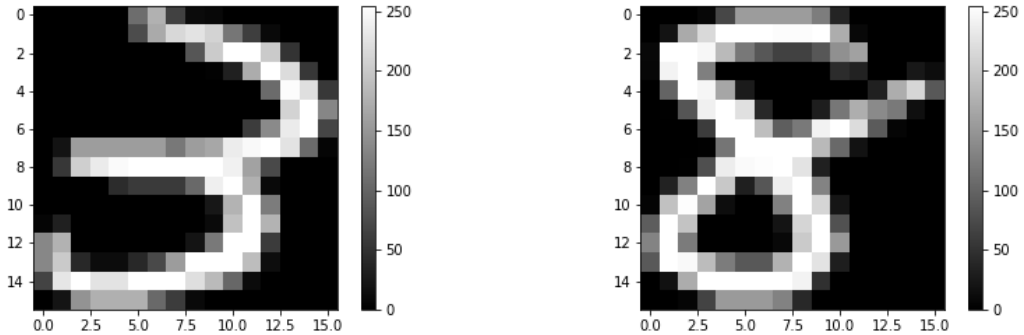
All z that is not in training set will be labeled as 0

Comment: One useful property of kernel functions is that the input space Z does not need to be a vector space; in other words, z does not need to be a feature vector. For all we know, Z can be turkeys in the world. As long as we can compute $k(z, z')$, kernel SVM works on turkeys.

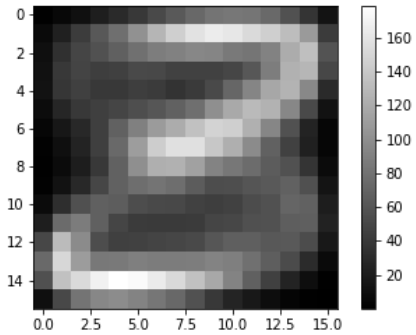
4 Principal Component Analysis [40 pts]

Download three.txt and eight.txt. Each has 200 handwritten digits. Each line is for a digit, vectorized from a 16x16 gray scale image.

1. (5 pts) Each line has 256 numbers: they are pixel values (0=black, 255=white) vectorized from the image as the first column (top down), the second column, and so on. Visualize the two gray scale images corresponding to the first line in three.txt and the first line in eight.txt.



2. (5 pts) Putting the two data files together (threes first, eights next) to form a $n \times D$ matrix X where $n = 400$ digits and $D = 256$ pixels. Note we use $n \times D$ size for X instead of $D \times n$ to be consistent with the convention in linear regression. The ith row of X is $x_i^\top$, where $x_i \in \mathbb{R}^D$ is the ith image in the combined data set. Compute the sample mean $y = \frac{1}{n} \sum_{i=1}^{n} x_i$. Visualize y as a 16x16 gray scale image.

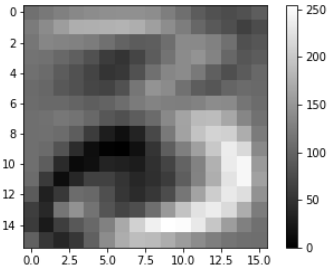


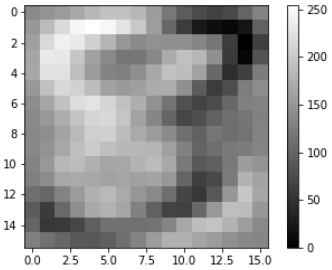
3. (10 pts) Center X using y above. Then form the sample covariance matrix $S = \frac{X^\top X}{n-1}$. Show the 5x5 submatrix $S(1 \ldots 5, 1 \ldots 5)$.

$$\begin{bmatrix} 59.17 & 142.15 & 28.68 & -7.18 & -14.34 \\ 142.15 & 878.94 & 374.14 & 24.13 & -87.13 \\ 28.68 & 374.14 & 1082.91 & 555.23 & 33.72 \\ -7.18 & 24.13 & 555.23 & 1181.24 & 777.77 \\ -14.34 & -87.13 & 33.72 & 777.77 & 1429.96 \end{bmatrix}$$

4. (10 pts) Use appropriate software to compute the two largest eigenvalues $\lambda_1 \geq \lambda_2$ and the corresponding eigenvectors v_1, v_2 of S. For example, in Matlab one can use eigs(S,2). Show the value of λ_1, λ_2. Visualize v_1, v_2 as two 16x16 gray scale images. Hint: their elements will not be in [0, 255], but you can shift and scale them appropriately. It is best if you can show an accompany "colorbar" that maps gray scale to values.

$\lambda_1 = 237155.2463$
$\lambda_2 = 145188.3527$

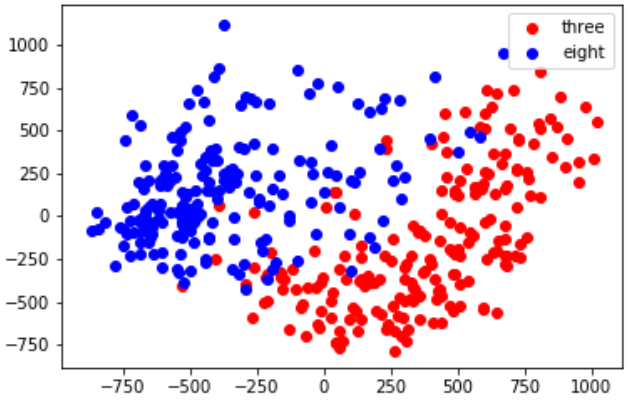




5. (5 pts) Now we project (the centered) X down to the two PCA directions. Let $V = [v_1 v_2]$ be the $D \times 2$ matrix. The projection is simply XV. Show the resulting two coordinates for the first line in three.txt and the first line in eight.txt, respectively.

first line in three: (136.21 , -242.63)
first line in eight: (-312.69 , 649.57)

6. (5 pts) Now plot the 2D point cloud of the 400 digits after projection. For visual interest, color points in three.txt red and points in eight.txt blue. But keep in mind that PCA is an unsupervised learning method and it does not know such class labels.


three
eight
1000
750
500
250
0
−250
−500
−750
−750
−500
−250
0
250
500
750
1000