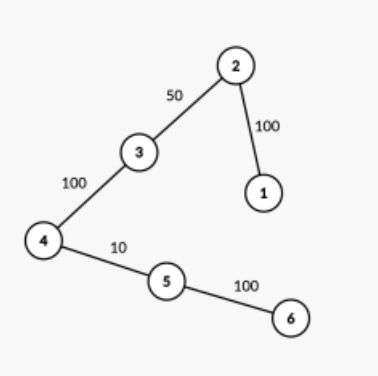
Worksheet on "Spectral Clustering, PCA and Linear Regression"

PRML – CS5691 (Jul–Nov 2023)

October 27, 2023

1. (5 marks) Consider the undirected, weighted graph given below. (a) Write down the Laplacian matrix **L** for this graph; (b) Write down the largest set of orthonormal eigen vectors of this graph Laplacian with eigen value 0, and (c) Redo part (b) for a graph where the edge (3,4) is removed.



Solution: (a) $\mathbf{L} = \begin{bmatrix} 100 & -100 & 0 & 0 & 0 & 0 \\ -100 & 150 & -50 & 0 & 0 & 0 \\ 0 & -50 & 150 & -100 & 0 & 0 \\ 0 & 0 & -100 & 110 & -10 & 0 \\ 0 & 0 & 0 & -10 & 110 & -100 \\ 0 & 0 & 0 & 0 & -100 & 100 \end{bmatrix}$

(b) The set comprising a single vector, which is the all ones vector divided by $\sqrt{6}$ is the solution. Eigen value 0 has multiplicity 1 (the number of connected components in the graph), so there is no other orthonormal set with more than one eigen vector for eigen value 0.

(c) Eigen value 0 has multiplicity 2. The two corresponding eigen vectors are the indicator vectors of the two components in the modified graph (normalized by the square root of the corresponding component's size), i.e., the vectors $\frac{1}{\sqrt{3}}[1\ 1\ 1\ 0\ 0\ 0]^T$, and $\frac{1}{\sqrt{3}}[0\ 0\ 0\ 1\ 1\ 1]^T$.

2. (5 marks) Consider the following data matrix, representing four sample points $x_n \in \mathbb{R}^2$

$$X = \begin{bmatrix} 4 & 1 \\ 2 & 3 \\ 5 & 4 \\ 1 & 0 \end{bmatrix}$$

Use principal component analysis (PCA) to represent the above data in only one direction. Report PC1 from the dataset, PC1-based representation of the last datapoint $x_4 = [1\ 0]^T$, and the reconstruction error of x_4.

Solution:

1. Mean $\mu = \begin{bmatrix} 3 & 2 \end{bmatrix}$
2. Mean centered data:

$$X' = \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 2 & 2 \\ -2 & -2 \end{bmatrix}$$

3. Covariance matrix:

$$(X')^T X' = \begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}$$

4. $\lambda = 16, 4$
5. Eigenvector corresponding to largest eigenvalue (16) is $PC1 = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]^T$.

   Project $x'_4 = [-2 \ -2]^T$ onto this PC1, and add mean vector back to find the PC1-based reconstruction/representation of x_4. Also calculate the resulting reconstruction error of this datapoint.

3. (6 marks) Consider the dataset of three datapoints below, and we would like to predict y using x.

$$\text{X} = \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 6 \end{bmatrix} \ \text{Y} = \begin{bmatrix} 3 \\ 6 \\ 7 \end{bmatrix}$$

   (a) Perform least squares regression and report the resulting regression coefficients w_{LS}.

   (b) Perform regularised least squares regression and find the optimal regression coefficients w_{RLS}, where the weight of the penalty term λ is assumed to be 1.

   (c) Assume a maximum likelihood approach for linear regression. Under this setting, **for a new datapoint** $x_{new} = [1 \ 1]^T$:

   (i) What is the predicted y_{new} value? Show your calculation.

   (ii) What is the uncertainty around y_{new} (in terms of the estimated variance of $y_{new}|x_{new}$)?

   (iii) Can you use this variance to not just report a single predicted value for y_{new}, but an interval that has 95% probability of containing the true y_{new} value? If so, specify this interval. (Note: Assume that the regression coefficients or its distribution estimated from the data are correct, and use the fact that approximately 95% of the values sampled from a normal distribution lie within two standard deviations from the mean.)

   (d) Answer the three sub-parts of part (3) above when Bayesian linear regression approach is used instead of MLE approach. Assume parameters α (precision parameter of Gaussian prior of w_i) and β (precision parameter of the Gaussian $y|x$) to each be 1.

Solution:

(a) $X^T X = \begin{bmatrix} 3 & 12 \\ 12 & 56 \end{bmatrix}$

$$(X^T X)^{-1} = \frac{1}{24} \begin{bmatrix} 56 & -12 \\ -12 & 3 \end{bmatrix}$$

$$w_{LS} = (X^T X)^{-1} X^T Y = \frac{1}{24} \begin{bmatrix} 32 \\ 24 \end{bmatrix}$$

(b) Repeat above calculation but with 1 added to each diagonal entry of $X^T X$.

(c) (i) Since $w_{LS} = w_{ML}$, we have $y_{new} = x_{new}^T w_{LS} = [1\ 1] w_{LS} = (32 + 24)/24$.

(ii) Use the formula for $(1/\hat{\beta}_{ML})$ from slides, which is basically the average of the squared residuals of the three datapoints, to calculate this variance.

(iii) Yes, the interval is $[\hat{\mu} - 2\hat{\sigma}, \hat{\mu} + 2\hat{\sigma}]$, where $\hat{\mu}$ is predicted y_{new} from part (a) and $\hat{\sigma}$ is $\sqrt{1/\hat{\beta}}$ from part (b) above. Substitute these values from the above parts and simplify the interval.

(d) The solution for Bayesian linear regression is similar, with the main difference being that the mean is calculated using w_{RLS} instead of w_{LS}, and the variance of the predictive distribution being a function of x_{new} instead of being the same for all x.

(i) Since $w_{RLS} = w_{MAP} = w_{mean-of-posterior}$, we have $y_{new} = x_{new}^T w_{RLS} = \ldots$ (complete calculation).

(ii) Refer to slides for the formulas involving parameters m_N, S_N for the posterior of w and results for the posterior predictive distribution. Verify that those formulas applied for the current problem yields: $var(y_{new}|x_{new}) = 1 + x_{new}^T S_N x_{new} = 1 + x_{new}^T (I + X^T X) x_{new}$. Complete the calculation.

(iii) Yes, the interval is $[\hat{\mu} - 2\hat{\sigma}, \hat{\mu} + 2\hat{\sigma}]$, where $\hat{\mu}$ is predicted y_{new} from part (a) and $\hat{\sigma}$ is the variance from part (b) above. Substitute these values from the above parts and simplify the interval.

Check how numerically different the Bayesian vs. MLE based linear regression answers above are.

4. (2 marks) Prove that a Laplacian matrix of an undirected simple graph is positive semi-definite.

Solution: We have already seen this in class using adjacency matrix representation A of a graph.

Another way to prove this is using the incidence matrix representation of a graph. If B is the incidence matrix of an orientation of the edges of G (each edge (i, j) in an undirected graph is only represented in one of the directions, either (i, j) or (j, i), but not both), then we can show $L = BB^T$. So $x^T L x = ||Bx||^2 \geq 0$.

Some students have asked for additional reference on Laplacian. One such reference is here: `https://www.cs.yale.edu/homes/spielman/561/2009/lect02-09.pdf`

5. (4 marks) We would like to reduce the dimensionality of a set of data points $D_N = \{x_n\}_{n=1}^N$ using PCA. Let u_i denote the ith PC of the dataset, and $\bar{x}$ the average of all the datapoints in D_N.

Let each $x_n \in \mathbb{R}^3$. Now, choose all formula(s) from below that will correctly compute the PC1-based reconstruction $\tilde{x}_n$ of x_n? Justify your answer.

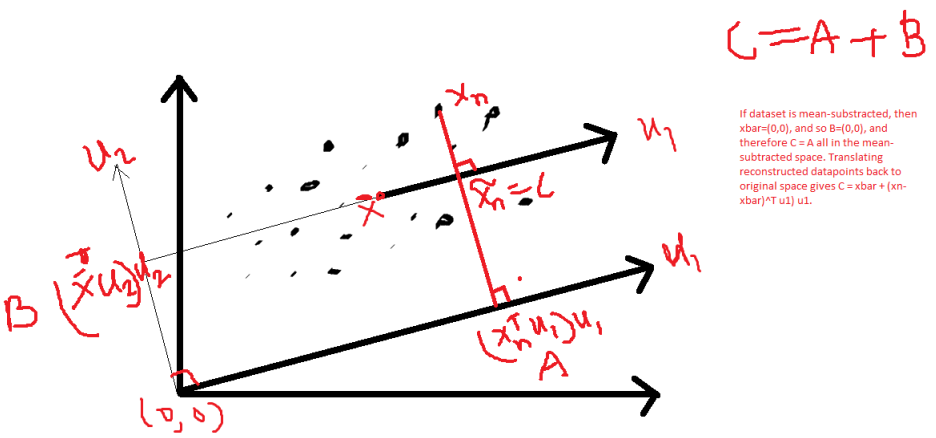
$$
\begin{aligned}
(F1): \quad \tilde{x}_n &= (x_n^T u_1)u_1 + (\bar{x}^T u_2)u_2 + (\bar{x}^T u_3)u_3 \\
(F2): \quad \tilde{x}_n &= \bar{x} + ((x_n - \bar{x})^T u_1)u_1 \\
(F3): \quad \tilde{x}_n &= (x_n^T u_1)u_1 + (x_n^T u_2)u_2 + (\bar{x}^T u_3)u_3 \\
(F4): \quad \tilde{x}_n &= (\bar{x}^T u_1)u_1 + (x_n^T u_2)u_2 + (x_n^T u_3)u_3
\end{aligned}
$$

Solution: F1 and F2. Both will give the same answer, because we've proved a more general statement in class below and we let $D = 3, M = 1$ in this proof to show that F1 and F2 above are the same.

If $x_n \in \mathbb{R}^D$ and $\tilde{x}_n$ is a reconstruction of x_n using only the top-M PCs of the dataset, then we've showed this proof in class:

$$\begin{aligned}
\tilde{x}_n &= \sum_{i=1}^{M}(x_n^T u_i)u_i + \sum_{i=M+1}^{D}(\bar{x}^T u_i)u_i \quad \text{(reconstn. formula from class slides)} \\
&= \bar{x} - \bar{x} + \sum_{i=1}^{M}(x_n^T u_i)u_i + \sum_{i=M+1}^{D}(\bar{x}^T u_i)u_i \quad \text{(add and subtract } \bar{x}\text{)} \\
&= \bar{x} - \sum_{i=1}^{D}(\bar{x}^T u_i)u_i + \sum_{i=1}^{M}(x_n^T u_i)u_i + \sum_{i=M+1}^{D}(\bar{x}^T u_i)u_i \quad \text{(represent } \bar{x} \text{ in terms of basis vectors } \{u_i\} \text{ of } \mathbb{R}^D\text{)} \\
&= \bar{x} + \sum_{i=1}^{M}((x_n - \bar{x})^T u_i)u_i
\end{aligned}$$

Geometrically also, you can try to visualize the above proof for D=2 and M=1 as below.



6. (2 marks) Suppose we have a data set with five predictors, X1 = GPA, X2 = IQ, X3 = Gender (1 for Female and 0 for Male), X4 = Interaction between GPA and IQ, and X5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$, $\hat{\beta}_5 = -10$.

   For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough. Is this statement correct? Justify.

Solution: False.

The least squares line is given by, $\hat{y}$ = 50 + 20 (GPA) + 0.07 (IQ) + 35 (Gender) + 0.01 (GPA × IQ) - 10 (GPA × Gender)

which becomes for the males, $\hat{y}$ = 50 + 20 (GPA) + 0.07 (IQ) + 0.01 (GPA × IQ),

and for the females it becomes, $\hat{y}$ = 85 + 10 (GPA) + 0.07 (IQ) + 0.01 (GPA × IQ).

So the starting salary for males is higher than for females on average if 50 + 20 (GPA) ≥ 85 + 10 (GPA), which is equivalent to GPA ≥ 3.5.

7. (4 marks) Consider the dataset D in the table below.

n	x (or x_n)	y (or y_n)
#1	1	1
#2	2	2
#3	4	3
#4	5	4
#5	6	4

(a) Find the line $y = wx + b$ that minimizes the squared vertical distance of the datapoints to the line (i.e., the squared errors in y). Specifically, find the w, b that minimizes

$$\sum_{n=1}^{5}((wx_n + b) - y_n)^2$$

.

(b) Find the line $y = mx + c$ that minimizes the squared perpendicular distance (i.e., shortest distance) of the datapoints to the line. Specifically, find the m, c that minimizes

$$\sum_{n=1}^{5}\frac{((mx_n + c) - y_n)^2}{m^2 + 1}$$

.

(Note: Distance of a point to a line from geometry is used to get each summation term above.)

(c) The two minimization problems above are each related to which ML task seen in class?

Solution: You can verify that linear regression least-squares solution (w_{LS} formula seen in class) will solve part (a), and PCA u_1 (PC1) formula will solve part (b). Use matrix notations and relevant matrix-vector-based formulas from class to obtain a quick solution to both parts. Doing the usual "setting the gradient to zero" approach to these minimization problems will also work, but may take longer time.

End Note: **Please also go through Assignment 2 questions and tutorials related to Spectral clustering, PCA, and Linear Regression.**