CS 181 Spring 2018 Section 9
Solution

1 Principal Component Analysis

1.1 Motivation

In many supervised learning problems, we try to find rich features that increase the expressivity of our model. In practice, this often involves using basis functions to transform our input into a higher dimensional space (eg. given data x, using x and x^2 as features, or using features learned by a neural network).

However, sometimes we want to reduce the dimensionality of our data. There can be several reasons: fewer features are easier to interpret (we might want to know why our model outputs a certain diagnosis, and only some of the patient record details will be relevant); models with fewer features are easier to handle computationally; and our data might be arbitrarily high-dimensional because of noise, so we would like to access the lower-dimensional *signal* from the data. One method for dimensionality reduction through **linear projections** of the original data is PCA. When reducing the dimensionality of our data from m to d, PCA can be interpreted as minimizing the reconstruction loss of projecting data onto a d basis vectors, or as maximizing the variance in data that can be explained by d basis vectors.

1.2 Finding the lower dimensional representation

To perform PCA, we first calculate the normalized **feature covariance** matrix:

$$\mathbf{S} = (\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top) = \mathbf{X}^\top \mathbf{X}$$

We then decide how many dimensions d out of the original m that we want to keep in the final representation (for visualizations, often this will be $d = 2$ or $d = 3$). We then find the d largest eigenvalues of $\mathbf{S}$. The $m \times 1$ eigenvectors $(\mathbf{u}_1, \ldots, \mathbf{u}_d)$ corresponding to these eigenvalues will be our lower-dimensional basis. Thus, we reduce the dimensionality of a data point $\mathbf{x}$ by projecting it onto this basis - we combine the eigenvectors into the $d \times m$ matrix $\mathbf{U}$, and compute $\langle \mathbf{x}^\top \mathbf{u}_1, \mathbf{x}^\top \mathbf{u}_2, \ldots, \mathbf{x}^\top \mathbf{u}_d, \ldots, 0 \rangle = \mathbf{U}\mathbf{x} = \mathbf{z}$. $\mathbf{z}$ is called the reconstruction coefficients where $\mathbf{U}^\top \mathbf{z}$ is the reconstruction of $\mathbf{x}$.

2 Bayesian Networks

A Bayesian network is a graphical model that represents random variables and their dependencies using a directed acyclic graph. Bayesian networks are useful because they allow us to efficiently model joint distributions over many variables by taking advantage of the local dependencies between variables. With Bayesian networks, we can easily reason about conditional independence and perform inference on large joint distributions.

2.1 D-separation rules

X_A and X_B are *d-separated* by evidence X_E if every undirected path from X_A to X_B is “blocked” by X_E. A path is blocked by evidence X_E if either:

1. There is a node Z with non-converging arrows on the path, and $Z \in X_E$.
2. There is a node Z with “converging arrows on the path, and neither Z nor its descendants are in X_E.

3 PCA

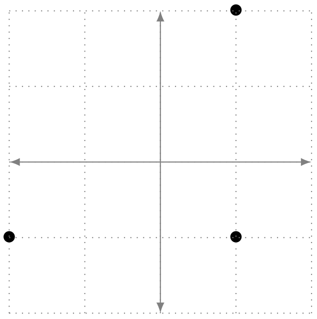
You are given the following data set:

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \mathbf{x}_3 = \begin{bmatrix} -2 \\ -1 \end{bmatrix}$$

You would like to use PCA to find a 1-dimensional representation of the data.

1. Plot the data set.
2. Compute the feature covariance matrix $\mathbf{S}$.
3. You find that $\mathbf{S}$ has eigenvector $[-1\ 1]^\top$ with eigenvalue 3 and eigenvector $[1\ 1]^\top$ with eigenvalue 9. What is the (normalized) basis vector $\mathbf{u}_1$ of your 1-dimensional representation? Add the basis vector $\mathbf{u}_1$ to your plot.
4. Compute the coefficients z_1, z_2, z_3. Add the lower-dimensional representations $z_1\mathbf{u}_1, z_2\mathbf{u}_1, z_3\mathbf{u}_1$ to your plot. Based on your plot, what is the relationship between $z_i\mathbf{u}_1$ and $\mathbf{x}_i$ with respect to the new basis?
5. Based on your plot, what would happen if you chose the unused eigenvector to be your basis vector?

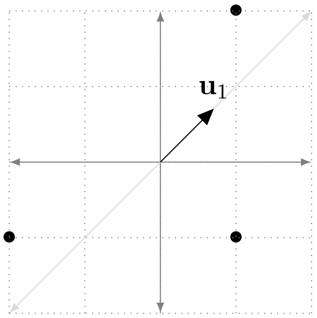
1.



2.

$$\mathbf{S} = \mathbf{X}^\top \mathbf{X} = \begin{bmatrix} 1 & -1 \\ 1 & 2 \\ -2 & -1 \end{bmatrix}^\top \begin{bmatrix} 1 & -1 \\ 1 & 2 \\ -2 & -1 \end{bmatrix} = \begin{bmatrix} 6 & 3 \\ 3 & 6 \end{bmatrix}$$

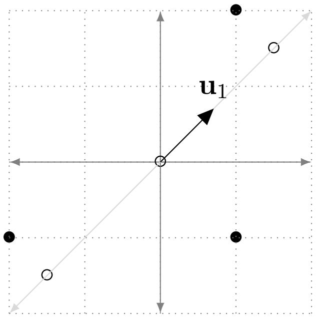
3. We select the eigenvectors with the largest eigenvalues for our basis, so our basis will contain a scalar multiple of $[1\ 1]^\top$. Normalizing $[1\ 1]^\top$ gives us that $\mathbf{u}_1 = [\frac{\sqrt{2}}{2}\ \frac{\sqrt{2}}{2}]^\top$.



4.

$$z_1 = \mathbf{x}_1^\top \mathbf{u}_1 = 0, \quad z_2 = \mathbf{x}_2^\top \mathbf{u}_1 = \frac{3\sqrt{2}}{2}, \quad z_3 = \mathbf{x}_3^\top \mathbf{u}_1 = -\frac{3\sqrt{2}}{2}$$

The open circles in the plot represent the lower-dimensional representation:



$z_i \mathbf{u}_1$ is the projection of $\mathbf{x}_i$ onto the basis vector.

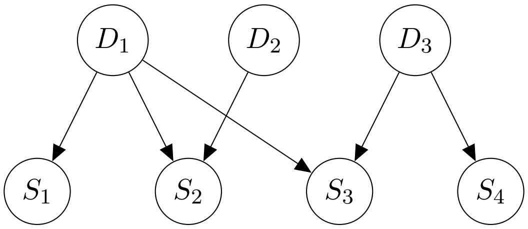
5. If we chose $[-1 \ 1]^\top$ to be the basis of our new representation, then the representation would capture less of the variance in the data. For example, $\mathbf{x}_2$ and $\mathbf{x}_3$ would be represented by the same point.

4 Network Basics

A patient goes to the doctor for a medical condition, and the doctor suspects 3 diseases as the cause of the condition. The 3 diseases are D_1, D_2, and D_3, and they are independent from each other (given no other observations). There are 4 symptoms S_1, S_2, S_3, and S_4, and the doctor wants to check for presence in order to find the most probable cause. S_1 can be caused by D_1, S_2 can be caused by D_1 and D_2, S_3 can be caused by D_1 and D_3, and S_4 can be caused by D_3. Assume all random variables are Bernoulli, i.e. the patient has the disease/symptom or not.

- **Q:** Draw a Bayesian network for this problem.

  A: Note that there are many valid networks (depending on the chosen variable ordering), some more efficient (i.e. requiring fewer parameters) than others. Here is a compact representation that comes from variable ordering $D_1, D_2, D_3, S_1, S_2, S_3, S_4$. (Recall that all dependencies to earlier variables need to be indicated with edges).



- **Q:** Write down the expression for the joint probability distribution given this network.

  A: $p(D_1, D_2, D_3, S_1, S_2, S_3, S_4)$

  $= p(D_1)p(D_2)p(D_3)p(S_1|D_1)p(S_2|D_1, D_2)p(S_3|D_1, D_3)p(S_4|D_3)$

- **Q:** How many parameters are required to describe this joint distribution?

  A:

Conditional Probability Table	Number of Parameters
$p(D_1)$	1
$p(D_2)$	1
$p(D_3)$	1
$p(S_1\|D_1)$	2
$p(S_2\|D_1, D_2)$	4
$p(S_3\|D_1, D_3)$	4
$p(S_4\|D_3)$	2
Total Number of Parameters	15

- **Q:** How many parameters would be required to represent the CPTs in a Bayesian network if there were no conditional independences between variables?

  A: The network would be structured as a clique, and considering order $D_1, D_2, D_3, S_1, S_2, S_3, S_4$, the number of parameters for the CPTs would be $1+2+4+8+16+32+64 = 127$. (We can see there is no saving relative to specifying the joint probability distribution directly, which would require $2^7 - 1 = 127$ numbers.)

- **Q:** What is an example of the 'explaining away' phenomenon in the compact Bayesian Network?

  A: S_3 depends on D_1 and D_3. When we know S_3, then conditioned on this D_1 and D_3 are not independent, and if we observe D_1 then D_3 is less likely to be a cause ("D_1 explains away D_3").

- **Q:** What diseases do we gain information about when observing the fourth symptom ($S_4 = \mathit{true}$)?

  A: We have independence relations $I(D_1, S_4)$ (since the path is blocked without observing S_3 and $I(D_2, S_4)$ (since the path is blocked at both S_2 and S_3). What is left is dependence between D_3 and S_4. Thus, we only learn information about D_3.

- **Q:** Suppose we know that the third symptom is present ($S_3 = \mathit{true}$). What does observing the fourth symptom ($S_4 = \mathit{true}$) tell us now?

  A: With $S_3 = \mathit{true}$, observing $S_4 = \mathit{true}$ now also gives us informaion about D_1 (via 'explaining away', or using d-separation, because the D_1 to S_4 path is no longer blocked at S_3). We still don't learn any information abhout D_2 because the D_2 to S_4 path remains blocked at S_2.

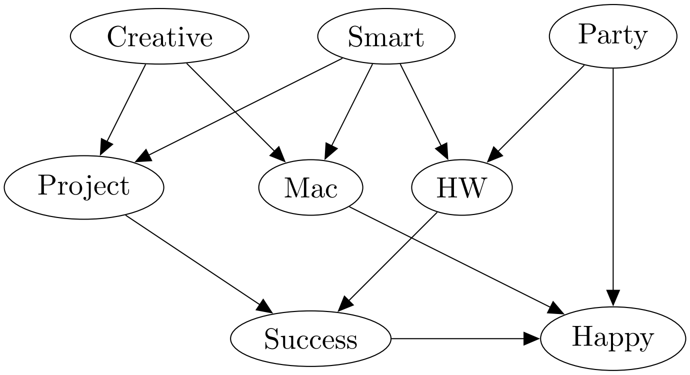
5 D-Separation

As part of a comprehensive study of the role of CS 181 on people's happiness, we have been collecting important data from students. In an entirely optional survey that all students are required to complete, we ask the following highly objective questions:

Do you party frequently [Party: Yes/No]?
Are you smart [Smart: Yes/No]?
Are you creative [Creative: Yes/No]? (Please only answer Yes or No)
Did you do well on all your homework assignments? [HW: Yes/No]
Do you use a Mac? [Mac: Yes/No]
Did your last major project succeed? [Project: Yes/No]
Did you succeed in your most important class? [Success: Yes/No]
Are you currently Happy? [Happy: Yes/No]

After consulting behavioral psychologists we build the following model:

- **Q:** True or False: *Party* is independent of *Success* given *HW*.



A: False; there is a path that is not blocked: $Party - HW - Smart - Project - Success$ has neither a converging arrows not in the set of evidence or a non-converging arrows in the set.

- **Q:** True or False: $Creative$ is independent of $Happy$ given Mac.

  A: False; there is a path that is not blocked: $Creative - Project - Success - Happy$

- **Q:** True or False: $Party$ is independent of $Smart$ given $Success$.

  A: False; there is a path that is not blocked between $Party$ and $Smart$: the path $Party - HW - Success$ is not blocked because the converging arrows node at HW has a descendant ($Success$) in the evidence.

- **Q:** True or False: $Party$ is independent of $Creative$ given $Happy$.

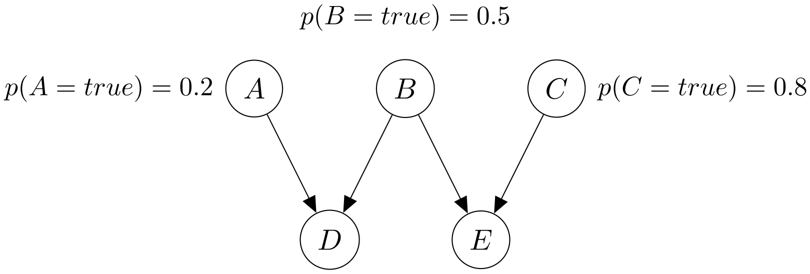
  A: False; there is a path that is not blocked between $Party$ and $Creative$ through the converging arrows at $Happy$. There are actually multiple not-blocked paths – can you find them?

- **Q:** True or False: $Party$ is independent of $Creative$ given $Success$, $Project$ and $Smart$.

  A: True! All paths between $Party$ and $Creative$ are blocked. Working from $Party$, the paths that come through $Happy$ are blocked there (converging arrows, no evidence). Those that come through HW and $Smart$ are blocked at $Smart$. Those that come through $HW, Success, Project$ are blocked at $Project$.

6 Inference

Consider the following Bayesian network, where all variables are Bernoulli.



A	B	$p(D = true \mid A, B)$
F	F	0.9
F	T	0.6
T	F	0.5
T	T	0.1

B	C	$p(E = true \mid B, C)$
F	F	0.2
F	T	0.4
T	F	0.8
T	T	0.3

- **Q:** What is the probability that all five variables are simultaneously $false$?

  A:

$$\begin{aligned} p(\neg A, \neg B, \neg C, \neg D, \neg E) &= p(\neg A)p(\neg B)p(\neg C)p(\neg D|\neg A, \neg B)p(\neg E|\neg B, \neg C) \\ &= (0.8)(0.5)(0.2)(0.1)(0.8) \\ &= 0.0064 \end{aligned}$$

- **Q:** What is the probability that A is $false$ given that the remaining variables are all known to be $true$?

  A: For this part, we need to calculate $p(\neg A|B, C, D, E)$.

  We know that $p(\neg A|B, C, D, E) \propto p(\neg A, B, C, D, E)$. The joint probabilities $p(\neg A, B, C, D, E)$ and $p(A, B, C, D, E)$ can be computed as:

$$\begin{aligned} p(\neg A, B, C, D, E) &= p(\neg A)p(B)p(C)p(D|\neg A, B)p(E|B, C) \\ &= (0.8)(0.5)(0.8)(0.6)(0.3) \\ &= (0.05760) \\ p(A, B, C, D, E) &= p(A)p(B)p(C)p(D|A, B)p(E|B, C) \\ &= (0.2)(0.5)(0.8)(0.1)(0.3) \\ &= (0.00240) \end{aligned}$$

  Finally, by normalization we have:

$$p(\neg A|B, C, D, E) = \frac{.05760}{.05760 + .00240} = .96$$