10-701/15-781, Fall 2006, Final

Dec 15, 5:30pm-8:30pm

- There are 9 questions in this exam (15 pages including this cover sheet).
- If you need more room to work out your answer to a question, use the back of the page and clearly mark on the front of the page if we are to look at what's on the back.
- This exam is open book and open notes. Computers, PDAs, cell phones are not allowed.
- You have 3 hours. Best luck!

Name:
Andrew ID:

Q	Topic	Max. Score	Score
1	Short Questions	20	
2	Instance-Based Learning	7	
3	Computational Learning Theory	9	
4	Gaussian Mixture Models	10	
5	Bayesian Networks	10	
6	Hidden Markov Models	12	
7	Dimensionality Reduction	8	
8	Graph-Theoretic Clustering	8	
9	MDPs and Reinforcement Learning	16	
	Total	100	

1 Short Questions (20pts, 2pts each)

(a) **True or [✓ False]**. The ID3 algorithm is guaranteed to find the optimal decision tree.

(b) **True or [✓ False]**. Consider a continuous probability distribution with density $f()$ that is nonzero everywhere. The probability of a value x is equal to $f(x)$.

(c) **[✓ True] or False**. In a Bayesian network, the inference results of the junction tree algorithm are the same as the inference results of variable elimination.

(d) **True or [✓ False]**. If two random variable X and Y are conditionally independent given another random variable Z, then in the corresponding Bayesian network, the nodes for X and Y are d-separated given Z.

(e) **[✓ True] or False**. Besides EM, gradient descent can be used to perform inference or learning on a Gaussian mixture model.

(f) In one sentence, characterize the differences between *maximum likelihood* and *maximum a posteriori* approaches.

maximum likelihood finds parameters to maximizes the likelihood function, while MAP maximizes the posterior probability

(g) In one sentence, characterize the differences between classification and regression.

classification maps inputs to discrete outputs
regression maps inputs to continuous outputs

(h) Give one similarity and one difference between feature selection and PCA.

similarity : reduce the dimension of data
difference : feature selection finds a subset of features, while PCA produces a smaller, new set

(i) Give one similarity and one difference between HMM and MDP.

similarity : Marov assumptions
difference : The Markov chain in HMM is hidden; in MDP, the states are fully observed

(j) For each of the following datasets, is it appropriate to use HMM? Provide a brief reasoning for your answer.

- [✓ Gene sequence dataset.]
- A database of movie reviews (eg., the IMDB database).
- [✓ Stock market price dataset.]
- [✓ Daily precipitation data from the Northwest of the US.]

Time-series data ; Markov assumption may be reasonable

2 Instance-Based Learning (7pts)

1. Consider the following training set in the 2-dimensional Euclidean space:

x	y	Class
-1	1	$-$
0	1	$+$
0	2	$-$
1	-1	$-$
1	0	$+$
1	2	$+$
2	2	$-$
2	3	$+$

Figure 1 shows a visualization of the data.

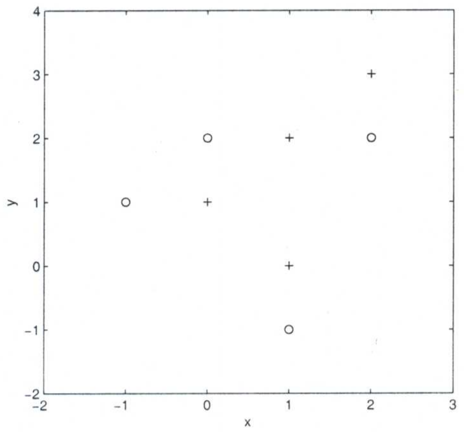


Figure 1: Dataset for Problem 2

(a) (1pt) What is the prediction of the 3-nearest-neighbor classifier at the point (1,1)?

+

(b) (1pt) What is the prediction of the 5-nearest-neighbor classifier at the point (1,1)?

+

(c) (1pt) What is the prediction of the 7-nearest-neighbor classifier at the point (1,1)?

−

2. Consider the two-class classification problem. At a data point x, the true conditional probability of a class $k, k \in \{0, 1\}$ is $p_k(x) = P(C = k|X = x)$.

   (a) (2pts) The Bayes error is the probability that an optimal Bayes classifier will misclassify a randomly drawn example. In terms of $p_k(x)$, what is the Bayes error E^* at x?

   $$1 - \max_{k \in \{0,1\}} P_k(x)$$

   or $$\min_{k \in \{0,1\}} P_k(x)$$

   (b) (2pts) In terms of $p_k(x)$ and $p_k(x')$ when x' is the nearest neighbor of x, what is the 1-nearest-neighbor error $E_{1\mathrm{NN}}$ at x?

   $$P_0(x)P_1(x') + P_0(x')P_1(x)$$

Note that asymptotically as the number of training examples grows, $E^* \leq E_{1\mathrm{NN}} \leq 2E^*$.

3 Computational Learning Theory (9pts, 3pts each)

In class we discussed different formula to provide a bound on the number of training examples sufficient for successful learning under different learning models.

$$m \geq \frac{1}{\epsilon}(\ln(1/\delta) + \ln|H|) \tag{1}$$

$$m \geq \frac{1}{2\epsilon^2}(\ln(1/\delta) + \ln|H|) \tag{2}$$

$$m \geq \frac{1}{\epsilon}(4\log_2(2/\delta) + 8VC(H)\log_2(13/\epsilon)) \tag{3}$$

Pick the appropriate one of the above formula to estimate the number of training examples needed for the following machine learning tasks. Briefly explain your choice.

1. Consider instances X containing 5 Boolean variables, $\{X_1, X_2, X_3, X_4, X_5\}$, and responses Y are $(X_1 \wedge X_4) \vee (X_2 \wedge X_3)$. We try to learn the function $f : X \rightarrow Y$ using a 2-layered neural network.

   (3). $|H| = \infty$, $Y \in H$.

2. Consider instances X containing 5 Boolean variables, $\{X_1, X_2, X_3, X_4, X_5\}$, and responses Y are $(X_1 \wedge X_4) \vee (X_2 \wedge X_3)$. We try to learn the function $f : X \rightarrow Y$ using a "depth-2 decision trees". A "depth-2 decision tree" is a tree with four leaves, all distance 2 from the root.

   (2). $|H| < \infty$, $Y \notin H$.

3. Consider instances X containing 5 Boolean variables, $\{X_1, X_2, X_3, X_4, X_5\}$, and responses Y are $(X_1 \wedge X_4) \vee (\neg X_1 \wedge X_3)$. We try to learn the function $f : X \rightarrow Y$ using a "depth-2 decision trees". A "depth-2 decision tree" is a tree with four leaves, all distance 2 from the root.

   (1). $|H| < \infty$, $Y \in H$.

4 Gaussian Mixture Model (10pts)

Consider the labeled training points in Figure 2, where '+' and 'o' denote positive and negative labels, respectively. Tom asks three students (Yifen, Fan and Indra) to fit Gaussian Mixture Models on this dataset.

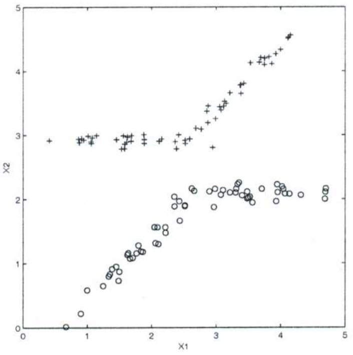
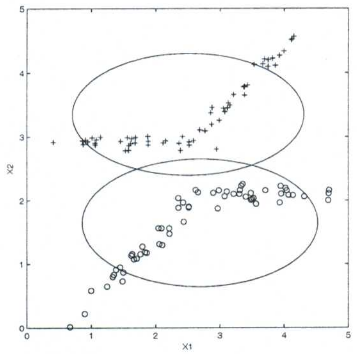
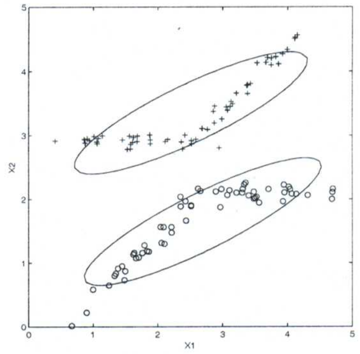


Figure 2: Dataset for Gaussian Mixture Model

1. (4pts) Yifen and Fan decide to use one Gaussian distribution for positive examples and one distribution for negative examples. The darker ellipse indicates the positive Gaussian distribution contour and the lighter ellipse indicates the negative Gaussian distribution contour.



Yifen's model



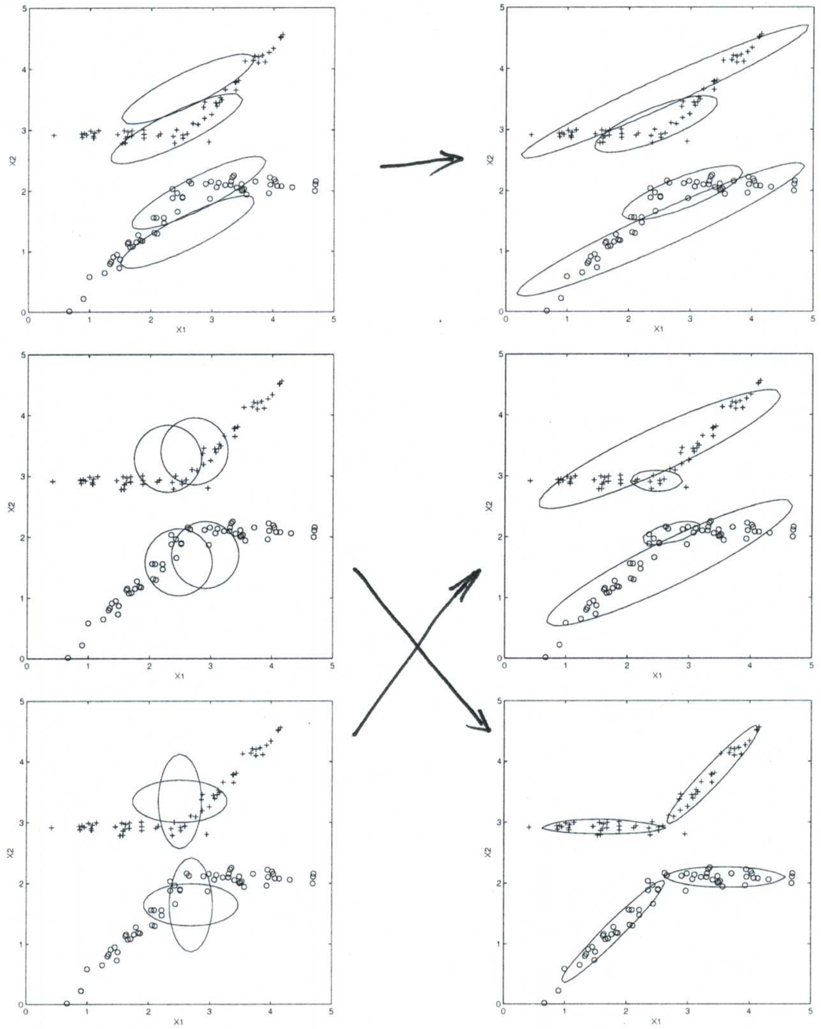
Fan's model

Whose model would you prefer for this dataset? What causes the difference between these two models?

Fan's model.

Yifen's model constrains the covariance matrixes to be diagonal, while Fan's model does not.

2. (6pts) Indra decides to use two Gaussian distributions for positive examples and two Gaussian distributions for negative examples. He uses EM algorithm to iteratively update parameters and also tries different initializations. The left column of Figure 3 shows 3 different initializations and the right column shows 3 possible models after the first iteration. For each initialization on the left, draw an arrow to the model on the right that will result after the first EM iteration. Your answer should consist of 3 arrows, one from each initialization.



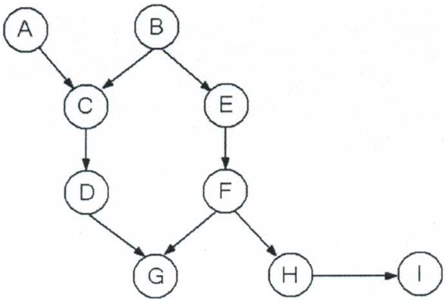
(a) Initialization

(b) After first iteration

Figure 3: Three different initializations and models after the first iteration.

5 Bayesian Networks (10pts)

The figure below shows a Bayesian network with 9 variables, all of which are binary.



1. (3pts) Which of the following statements are always true for this Bayes net?

   (a) $P(A, B|G) = P(A|G)P(B|G)$;
   [✓ (b)] $P(A, I) = P(A)P(I)$;
   (c) $P(B, H|E, G) = P(B|E, G)P(H|E, G)$;
   (d) $P(C|B, F) = P(C|F)$.

2. (2pts) What is the number of independent parameters in this graphical model?

   20

3. (3pts) The computational complexity of a graph elimination algorithm is determined by the size of the maximal elimination clique produced in the elimination process. What is the minimum size of such maximal elimination clique when we choose a perfect elimination order to compute $P(C = 1)$ using the graph elimination algorithm?

   3

4. (2pts) We would like to compute

$$\mu = \frac{P(F = 1|A, B, C, D, E, G, H, I)}{P(F = 0|A, B, C, D, E, G, H, I)}$$

   The value of μ depends on the values of all the variables other than F. What is the maximum possible number of different values of μ?

   16

*Given the value of μ, as in the setting of Gibbs sampling, we could draw the random variable F from a Bernoulli distribution: $F \sim \text{Bernoulli}[1/(1 + \mu^{-1})]$.

6 Hidden Markov Models (12pts)

Consider an HMM with states $Y_t \in \{S_1, S_2, S_3\}$, observations $X_t \in \{A, B, C\}$, and parameters

$\pi_1 = 1$	$a_{11} = 1/2$	$a_{12} = 1/4$	$a_{13} = 1/4$	$b_1(A) = 1/2$	$b_1(B) = 1/2$	$b_1(C) = 0$
$\pi_2 = 0$	$a_{21} = 0$	$a_{22} = 1/2$	$a_{23} = 1/2$	$b_2(A) = 1/2$	$b_2(B) = 0$	$b_2(C) = 1/2$
$\pi_3 = 0$	$a_{31} = 0$	$a_{32} = 0$	$a_{33} = 1$	$b_3(A) = 0$	$b_3(B) = 1/2$	$b_3(C) = 1/2$

(a) (3pts) What is $P(Y_5 = S_3)$?

$$1 - P(Y_5 = S_1) - P(Y_5 = S_2)$$
$$= 1 - \frac{1}{16} - 4 \times \frac{1}{32}$$
$$= \frac{13}{16}$$

For 6(b)-(d), suppose we observe $AABCABC$, starting at time point 1.

(b) (2pts) What is $P(Y_5 = S_3 | X_{1:7} = AABCABC)$?

0

(c) (4pts) Fill in the following table assuming the observation $AABCABC$. The α's are values obtained during the forward algorithm: $\alpha_t(i) = P(X_1, \ldots, X_t, Y_t = i)$.

t	$\alpha_t(1)$	$\alpha_t(2)$	$\alpha_t(3)$
1	$\frac{1}{2}$	0	0
2	$\frac{1}{8}$	$\frac{1}{16}$	0
3	$\frac{1}{32}$	0	$\frac{1}{32}$
4	0	$\frac{1}{2^8}$	$\frac{5}{2^8}$
5	0	$\frac{1}{2^{10}}$	0
6	0	0	$\frac{1}{2^{12}}$
7	0	0	$\frac{1}{2^{13}}$

(d) (3pts) Write down the sequence of $Y_{1:7}$ with the maximal posterior probability assuming the observation $AABCABC$. What is that posterior probability?

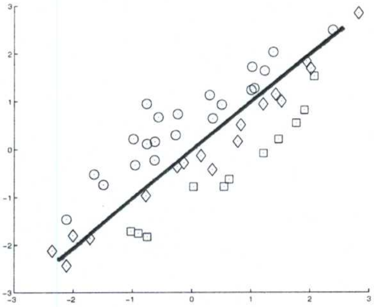
$S_1 S_1 S_1 S_2 S_2 S_3 S_3$

posterior probability : 1

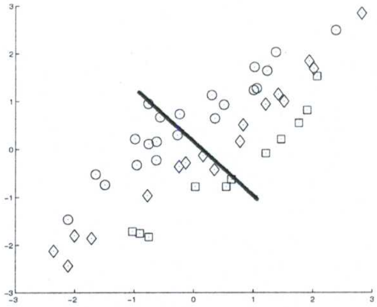
7 Dimensionality Reduction (8pts)

In this problem four linear dimensionality reduction methods will be discussed. They are principal component analysis (PCA), linear discriminant analysis (LDA), canonical correlation analysis (CCA), non-negative matrix factorization (NMF).

1. (3pts) LDA reduces the dimensionality given labels by *maximizing the overall interclass variance relative to intraclass variance.* Plot the directions of the *first* PCA and LDA components in the following figures respectively.

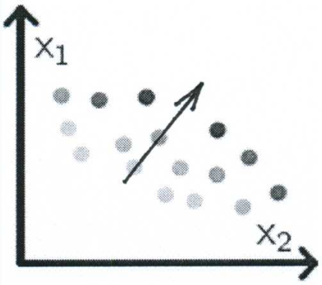


1(a) First PCA component

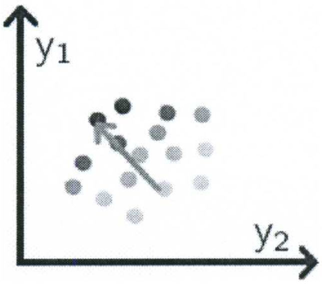


1(b) First LDA component

2. (2pts) In practice, each data point may have multiple vector-valued properties, e.g. a gene has its expression levels as well as the position on the genome. The goal of CCA is to reduce the dimensionality of the properties jointly. Suppose we have data points with two properties $\mathbf{x}$ and $\mathbf{y}$, each of which is a 2-dimension vector. This 4-dimensional data is shown in the pair of figures below; different data points are shown in different gray scales. CCA *finds* $(\mathbf{u}, \mathbf{v})$ *to maximize the correlation* $\widehat{corr}(\mathbf{u}^T\mathbf{x})(\mathbf{v}^T\mathbf{y})$. In figure 2(b) we have given the direction of vector $\mathbf{v}$, plot the vector $\mathbf{u}$ in figure 2(a).

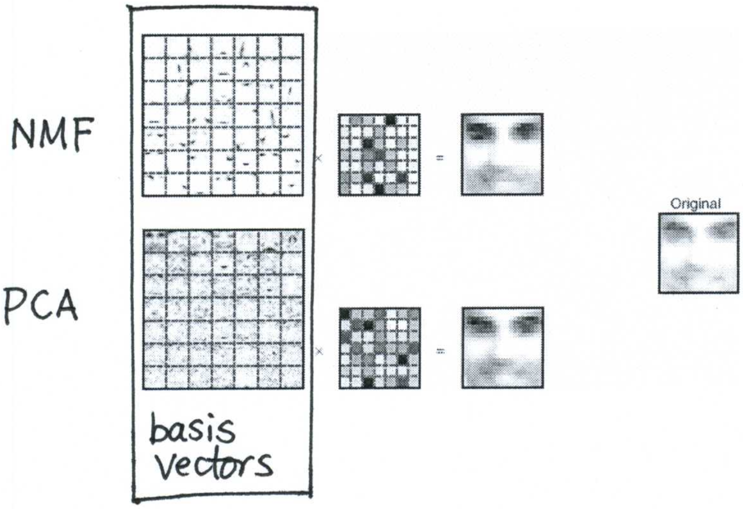


2(a)



2(b)

3. (3pts) The goal of NMF is to reduce the dimensionality given non-negativity constraints. That is, we would like to find principle components $\mathbf{u}_1, \ldots, \mathbf{u}_r$, each of which is of dimension $d > r$, such that the d-dimensional data $\mathbf{x} \approx \sum_{i=1}^{r} z_i \mathbf{u}_i$, and all entries in $\mathbf{x}, \mathbf{z}, \mathbf{u}_{1:r}$ are non-negative. NMF tends to find sparse (usually small L1 norm) basis vectors $\mathbf{u}_i$'s . Below is an example of applying PCA and NMF on a face image. Please point out the basis vectors in the equations and give them correct labels (NMF or PCA).



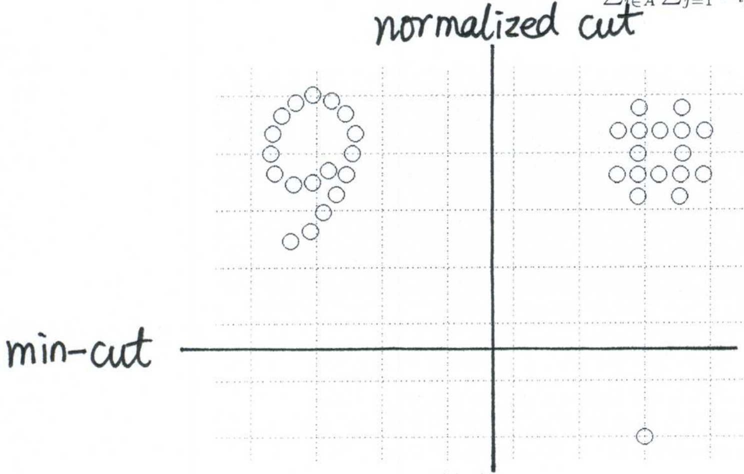
(* Figures in 7-2, 7-3 are originally from `http://www.eecs.berkeley.edu/~asimma/294-fall06/lectures/dimension/talk-maximal-1x2.pdf`.)

8 Graph-Theoretic Clustering (8pts)

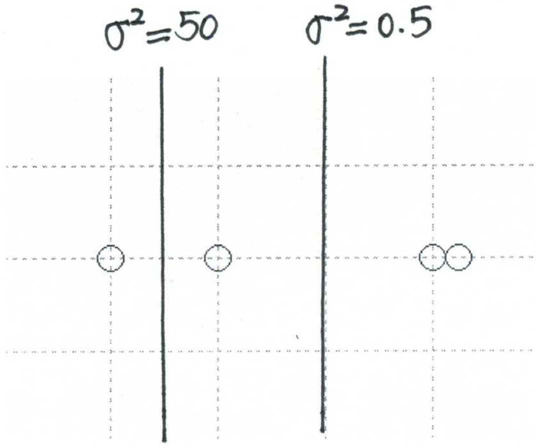
Part A. Min-Cut and Normalized Cut

In this problem, we consider the 2-clustering problem, in which we have N data points $\mathbf{x}_{1:N}$ to be grouped in two clusters, denoted by A and B. Given the N by N affinity matrix W,

- Min-Cut: minimizes $\sum_{i\in A}\sum_{j\in B} W_{ij}$;
- Normalized Cut: minimizes $\frac{\sum_{i\in A}\sum_{j\in B} W_{ij}}{\sum_{i\in A}\sum_{j=1}^{N} W_{ij}} + \frac{\sum_{i\in A}\sum_{j\in B} W_{ij}}{\sum_{i=1}^{N}\sum_{j\in B} W_{ij}}$.



(A1)



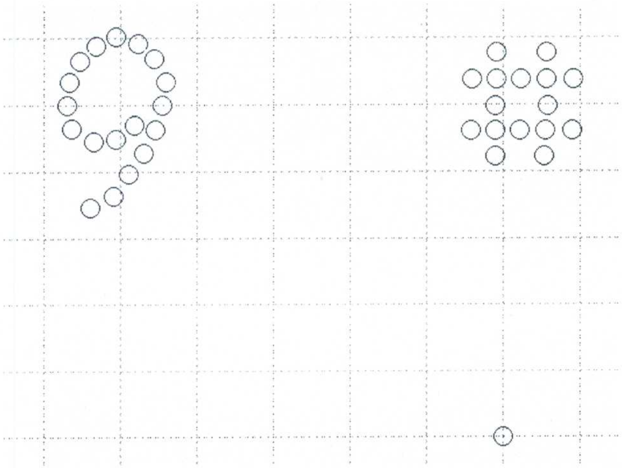
(A2)

A1. (2pts) The data points are shown in Figure (A1) above. The grid unit is 1. Let $W_{ij} = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}$, give the clustering results of min-cut and normalized cut respectively (You may show your work in the figure directly).

A2. (2pts) The data points are shown in Figure (A2) above. The grid unit is 1. Let $W_{ij} = e^{\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}}$, describe the clustering results of min-cut algorithm for $\sigma^2 = 50$ and $\sigma^2 = 0.5$ respectively.

Part B. Spectral Clustering

Now back to the setting of the 2-clustering problem A1. The grid unit is 1.



B1. (2pts) If we use Euclidean distance to construct the affinity matrix W as follows:

$$W_{ij} = \begin{cases} 1 & \text{if } \| \mathbf{x}_i - \mathbf{x}_j \|_2^2 \leq \sigma^2 \\ 0 & \text{otherwise} \end{cases}$$

What σ^2 value would you choose? Briefly explain.

$\sigma^2 = 9 \sim 16$. W_{ij} should be 1 for every pair of points within "9", "#"; it should be 0 for other cases.

B2. (2pts) The next step is to compute the $k = 2$ dominant eigenvectors of the affinity matrix W. For the value of σ^2 you chose in the previous question, can you compute analytically eigenvalues corresponding to the first two eigenvectors? If yes, compute and report the eigenvalues. If not, briefly explain.

$$W = \begin{bmatrix} 1_{18\times18} & 0 & 0 \\ 0 & 1_{16\times16} & \\ 0 & 0 & 1 \end{bmatrix}$$

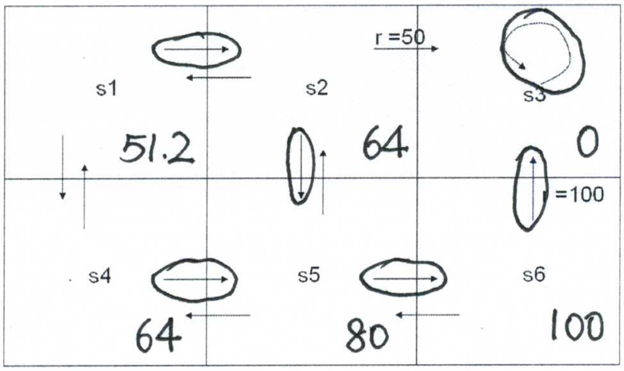
First two eigenvalues : 18, 16.

B3. *(1 Extra Credit, please try this question after you finished others!)
Suppose the data is of very high dimension so that it is impossible to visualize them and pick a good value as we did in Part B1. Suggest a heuristic that could find an appropriate σ^2.

9 MDPs and Reinforcement Learning [16pts]

Part A. [10pts]

Consider the following deterministic Markov Decision Process (MDP), describing a simple robot grid world. Notice the values of the *immediate rewards* are written next to transitions. Transitions with no value have an immediate reward of 0. **Assume the discount factor** $\gamma = 0.8$.



A1. (2pts) For each state s, write the value for $V^*(s)$ inside the corresponding square in the diagram.

A2. (2pts) Mark the state-action transition arrows that correspond to one *optimal* policy. If there is a tie, always choose the state with the smallest index.

A3. (2pts) Give a different value for γ which results in a different optimal policy and the number of changed policy actions should be minimal. Give your new value for γ, and describe the resulting policy by indicating which $\pi(s)$ values (i.e., which policy actions) change.

New value for γ: 0.7

Changed policy actions: $\pi(S_2) = S_3$

For the remainder of this question, assume again that $\gamma = 0.8$.

A4. (2pts) How many complete loops (iterations) of value iteration are sufficient to guarantee finding the optimal policy for this MDP? Assume that values are initialized to zero, and that states are considered in an arbitrary order on each iteration.

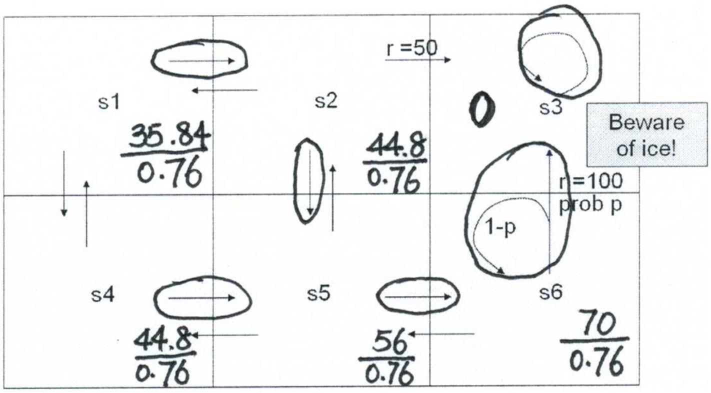
4

A5. (2pts) Is it possible to change the immediate reward function so that V^* changes but the optimal policy π^* remains unchanged? If yes, give such a change, and describe the resulting change to V^*. Otherwise, explain in at most 2 sentences why this is impossible.

Yes. Double each immediate reward. Then V^* is also doubled; π^* remains unchanged.

Part B. (6pts)

It is December. Unfortunately for our robot, a patch of ice has appeared in its world, making one of its actions non-deterministic. The resulting MDP is shown below. Note that now the result of the action "go north" from state $s6$ results in one of two outcomes. With probability p the robot succeeds in transitioning to state $s3$ and receives immediate reward 100. However, with probability $(1 - p)$ it slips on the ice, and remains in state $s6$ with zero immediate reward. **Assume the discount factor** $\gamma = 0.8$.



B1. (4pts) Assume $p = 0.7$. Write in the values of V^* for each state, and circle the actions in the optimal policy.

$$V_6^* = 100p + \gamma(1-p)V_6^*$$

B2. (2pts) How bad does the ice have to get before the robot will prefer to completely avoid it? Answer this question by giving a value for p below which the optimal policy chooses actions that completely avoid the ice, even choosing the action "go west" over "go north" when the robot is in state s6.

$$50\gamma^2 = V_6^* = \frac{100p}{1-\gamma(1-p)}$$

$$p = \frac{\gamma^2-\gamma^3}{2-\gamma^3} = \frac{8}{93}$$