10-601 Machine Learning, Final Exam: Fall 2013

- Please put your name on this cover sheet

- You are only allowed one A4 size cheat sheet,no electronics, no applets, no wireless communication

- **There are 114 points total on the entire exam, and you have 80 minutes to complete it. Good luck!**

Name:

Andrew ID:

Q	Topic	Max. Score	Score
1	True/False	16	
2	Multiple choice	8	
3	Short Answer	32	
4	Interpreting Data	10	
5	Linear Classifier	10	
6	Decision Tree	4	
7	SVM	6	
8	Compare Classifiers	6	
9	Ensemble and semi-supervised learning	10	
10	LDA	12	
	Total	114	

1 True/False [16pts]

Use one or two sentences to explain your answer.

1. [2pts] Linear regression can only deal with linear relationships.

2. [2pts] Theoretically the normal equation method and the least-mean-square (LMS) method should generate the same results for linear regression problem.

3. [2pts] The dual form of SVM solves the exact same problem as primary using a different representation.

4. [2pts] The Dirichlet distribution is a conjugate prior for the multinomial distribution.

5. [2pts] It is impossible to have infinite VC dimension for any set of functions.

6. [2pts] When a dataset is not linearly separable, no SVM classifiers will find a hyperplane that separates the data points perfectly.

7. [2pts] Logistic Regression is a linear classifier.

8. [2pts] For a given fixed set of data points and a fixed k, k-means always converges to the same clustering of the data.

2   Multiple Choices [8pts]

Choose one or more correct answers in the following questions.

2.1   In the k-nearest neighbor classifier, which of the following statements are true? [2pts]

- A kNN is a supervised classifier.

- B The hyperparameter k in kNN is typically set to an odd number.

- C When k is set to an extremly large number, it is more likely that the classifier will overfit than underfit.

- D Both kNN and k-means are unsupervised learning techniques.

2.2   Which of the following statements about stochastic gradient descent (SGD) are true? [2pts]

- A If an objective function has only one global optimum, then SGD is guaranteed to converge to the global optimum.

- B The standard SGD (excluding the mini-batch variant) is an online algorithm that updates the weights by looking at one example at a time.

- C In practice, when we set an inappropriate learning rate in SGD, it might take longer time for SGD to converge.

- D SGD can also be used to train a conditional random field (CRF) model.

2.3   Which of the following statements about the hidden Markov model (HMM) are true? [2pts]

- A The forward-backward algorithm is a standard inference algorithm for HMM.

- B Viterbi algorithm is a sum product algorithm for finding the most likely sequence of hidden states in a HMM.

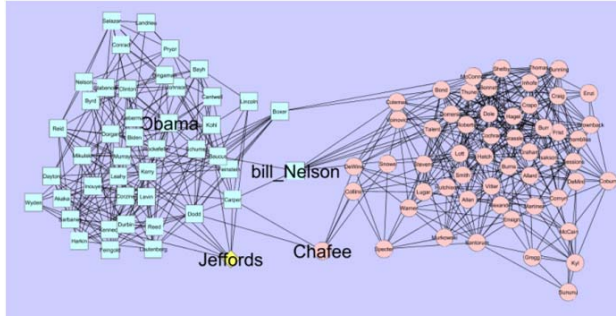- C HMM is a generative sequence modeling technique.

Figure 1: Network data.

2.4 Which of the following models would be possible for use with the data shown in figure 1 given that you have no labels for the nodes? [2pts]

The data shows the voting record for 100 senator in march 2005.

- A k-means

- B Naive Bayes

- C Stochastic block models (SBM) with 100 blocks, one for each senator.

- D mixed membership SBM with 100 blocks, one for each senator.

- E SBM with 2 blocks, one for each party

- F mixed-membership SBM with 2 blocks, one for each party.

3 Short answers [32pts]

1. [2pts] Given samples from a uni-variate Gaussian distribution with mean 0 and variance 1, how could you create samples from a bi-variate Gaussian with means μ_1, μ_2 and co-variance

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}$$

2. [2pts] Suppose that in answering a question in a multiple choice test, an examinee either knows the answer, with probability p, or he guesses the answer with probability $1 - p$. Assume that the probability of answering a question correctly is $1 - \delta$ for an examine who knows the answer and $1/m$ for an examinee who guesses, where m is the number of multiple choice alternatives. What is the probability that an examinee knew the answer to a question given that he correctly answered it.

3. [2pts] Why do we need smoothing in the naive Bayes classifier?

4. [2pts] What is the main difference between semi-supervised learning and unsupervised learning?

5. [2pts] Figure 2 shows the true distribution of the two classes. If your classifier discovers the true distribution then draw the decision boundary for your classifier. Also state whether this best classifier will have zero error or not.
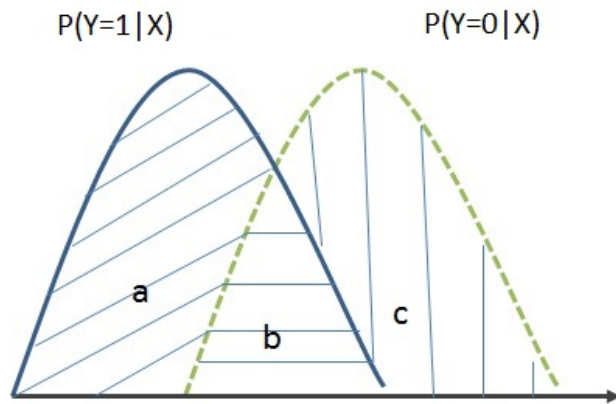
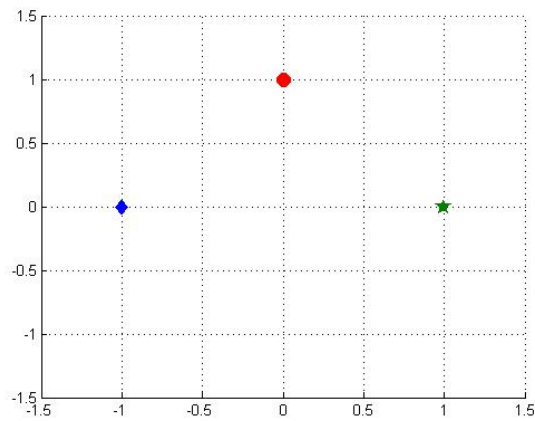Figure 2: Probability density of the two classes.

Figure 3: Points on two dimentional plane.

6. [2pts] For the data given in figure 3 draw the 1-NN decision boundary. Different shape represents Different classes.

7. [2pts] What is the VC-dimension of the class of circles centered at the origin origin in $\mathcal{R}^2$, or in other words the set of fucntions defined as

$$f(x_1, x_2) = \begin{cases} 1 & \text{if } (x_1)^2 + (x_2)^2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

8. [2pts] Suppose you have regression data generated from a polynomial of degree 5. Characterize the bias-variance of the estimates of the following models on the data with respect to the true model by circling appropriate entries

	Bias	Variance
Linear Regression	low/high	low/high
Polynomial of degree 5	low/high	low/high
Polynomial of degree 20	low/high	low/high

9. [2pts] When optimizing SVM with slack variables, we have the following objective function:

$$\arg\min_{w,z} \frac{1}{2}||\vec{w}||^2 + C\sum_i z_i \tag{1}$$

$$s.t. \forall i, y_i \vec{w} \cdot \vec{x}_i + z_i > 1, \ z_i \geq 0 \tag{2}$$

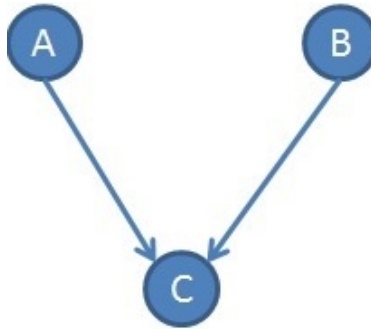When we change C, how does the size of the margin change?

Figure 4: directed graphical model

10. [2pts] What are the independencies denoted by the graph in figure 4? Can the independencies represented by this directed graph be represented by an undirected graphical model?
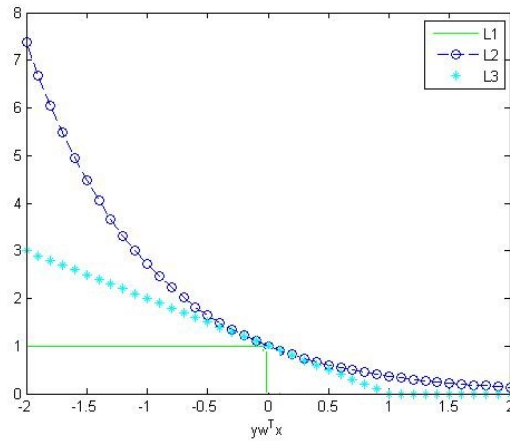
Figure 5: Loss function

11. [2pts] Label $L1$, $L2$ and $L3$ in figure 5 as one of "square loss" (used in regression), "exponential loss" (example used in adaboost), "zero one error" and "hinge loss" (example used in SVM).

12. [2pts] Name one method to enable linear regression to be used for modeling non-linear relationships.

13. [2pts] What is the reason for using regularization methods in machine learning problems?

14. [2pts] What property does the l_1 regularization lead to?

15. [2pts] Show that
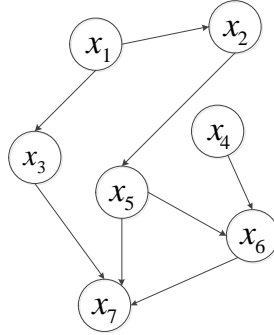
$$p(A|B,C) = \frac{P(A,B|C)}{P(B|C)} \qquad (3)$$

Figure 6: Graphical Model

16. [2pts] We have seven random variables $x_1, x_2, \cdots, x_7$ and define a Bayesian network over them drawn in figure 6. From the Bayesian network, write down the joint distribution $p(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$ of the seven variables.
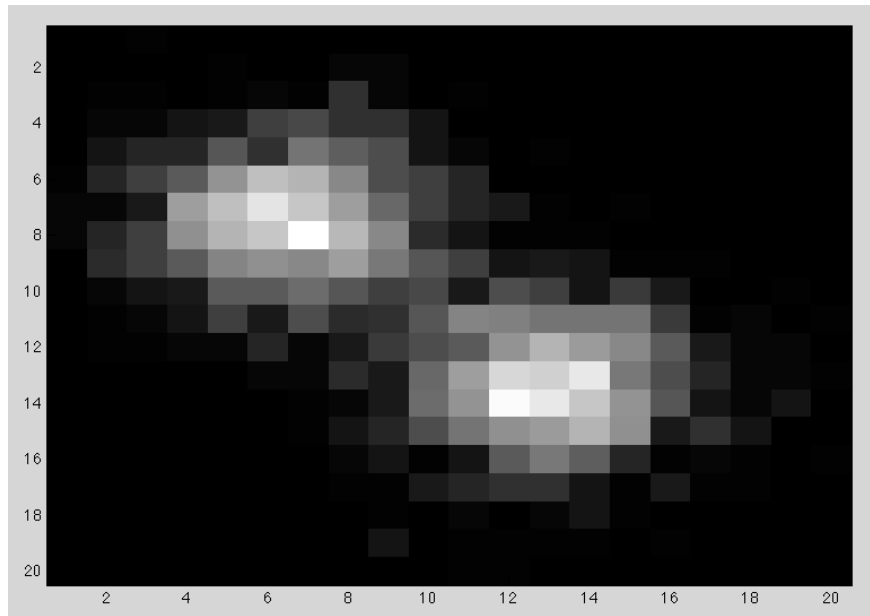
4   Interpreting data [10pts]

Figure 7: Summaries of a 2-dimensional dataset

Figure 7 and figure 8 shows some Matlab images discussed in a lecture. These images summarize a data sample S where each instance is a point $\mathbf{x} = (x_1, x_2)$ in two-dimensional space. Figure 7 is a histogram where each dimension is placed into one of 20 equal-sized bins (white is a high count, black is a low count) and Figure 8 is a scaled grey image `imagesc` of the actual values of the data.

1. [2pts] About how many instances are in S?

2. [2pts] Which of the following methods would be *reasonable* to use with this data?

   (a) k-means

   (b) Gaussian naive Bayes

   (c) k-nearest-neighbor

   (d) mixtures of Gaussians

   (e) logistic regression

   (f) principal component analysis

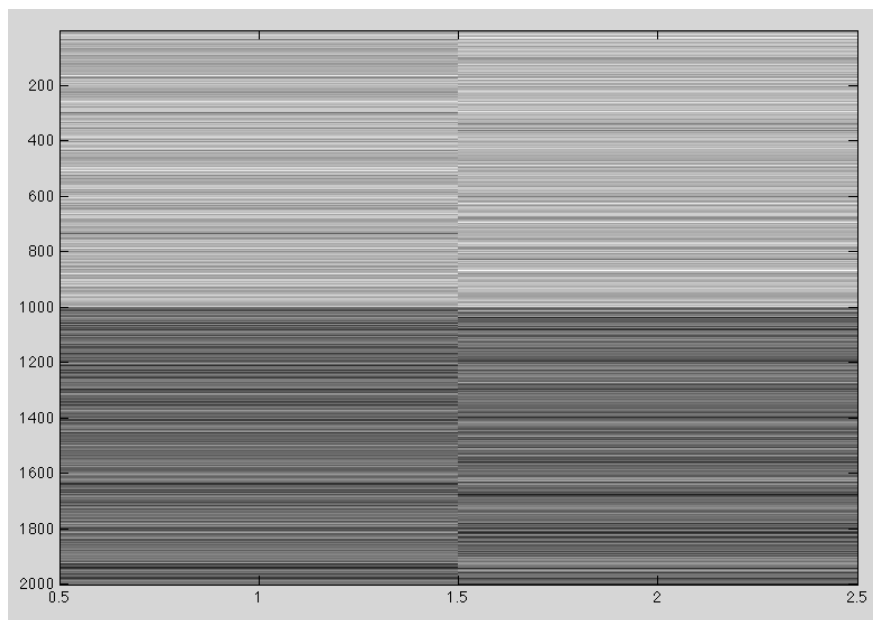3. [2pts] Which of the methods above would be *most appropriate* to use with this data, and why?

Figure 8: Summaries of a 2-dimensional dataset

4. [2pts] Let X_1 and X_2 be random variables corresponding to the two dimensions of the data. From this sample, do you think that X_1 and X_2 are independent? Why or why not?

5. [2pts] Suppose S is a sample of the *positive* examples of a labeled data: i.e., all the instances in S have label $+1$. Kevin believes that the best classifier learner to use for this task is Gaussian naive Bayes. Do you agree? why or why not?

5   Linear classifiers [10pts]

1. [2pts] Kevin wants to improve the accuracy of his implementation of a multinomial naive Bayes text classifier by using a new smoothing method. How would he be most likely to improve performance, and why?

   (a) By smoothing the estimates for the class priors.
   (b) By smoothing the estimates for the conditional probabilities of features given a class.

2. [2pts] Consider the perceptron applied to a series of examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, where k is the number of mistakes, $\mathbf{v}_j$ is the j-th classifier produced by the perceptron, and R^2 is an upper bound on $\mathbf{x}_i \cdot \mathbf{x}_i$ for all instances $\mathbf{x}_i$. Suppose also that the data is separable with margin γ, as defined in the lecture.

   Which of the following must be true?

   (a) There is some vector $\mathbf{u}$ such that for all i, $(\mathbf{u} \cdot x_i)y_i > \gamma$ and $\mathbf{u}$ is orthogonal to $\mathbf{x}_i$.
   (b) The final classifier $\mathbf{v}_k$ produced by the perceptron will have an error rate of no more than k/n on the training examples.
   (c) $k < \frac{R^2}{\gamma^2}$
   (d) $k < \frac{\gamma^2}{R^2}$
   (e) $\gamma \geq 1$
   (f) $\gamma \geq 0$

3. [2pts] Suppose multinomial naive Bayes gets zero *training* error on a binary dataset D. Does that imply that D is linearly separable? Justify your answer.

4. A common regularization term for logistic regression is

$$\mu \sum_{j=1}^{d} (w_j)^2$$

   where d is the number of dimensions of the data, and w_j is the weight for the j-th parameter.

(a) [2pts] In Kevin's experiments with logistic regression, he used cross-validation to pick the best value of μ among these values: -1, -0.5, -0.1, 0, 0.1, 0.5, 1. He asks your advice about this selection—which set of values will you ask him to use among his suggested values and why?

(b) [2pts] Kevin wants to explore using a regularization term of the form $\mu \sum_{j=1}^{d} (w_j)^k$ for $k = 1, 2, 3, 4$. He asks your advice on this proposal—what do you say?

6 Decision Tree [4pts]

Assume there are 4 binary attributes (value can only be T or F): Attribute#1, Attribute#2, Attribute#3, Attribute#4, two kinds of labels: 0 and 1. Now you are given the following 8 instances

Table 1: Instances

Instance Label	Attribute#1	Attribute#2	Attribute#3	Attribute#4
1	T	T	T	F
1	T	T	T	F
1	F	T	T	F
1	F	T	F	F
0	T	T	F	F
0	T	T	F	F
0	F	T	F	F
0	F	T	T	F

1 [2pts] Compute the information gain for selecting each attribute as the separating attribute in the first (root) node. Which attribute should be selected?

2 [2pts] Does there exist a decision tree which can classify the given instances perfectly? If yes, draw that decision tree, otherwise give a simple explanation.

7 Support Vector Machine [6pts]

1. [2pts] Given the data points shown in the following figure with two different labels. Draw the linear SVM decision boundary for this binary classification problem. Also circle the support vectors (data points) for the boundary you find out
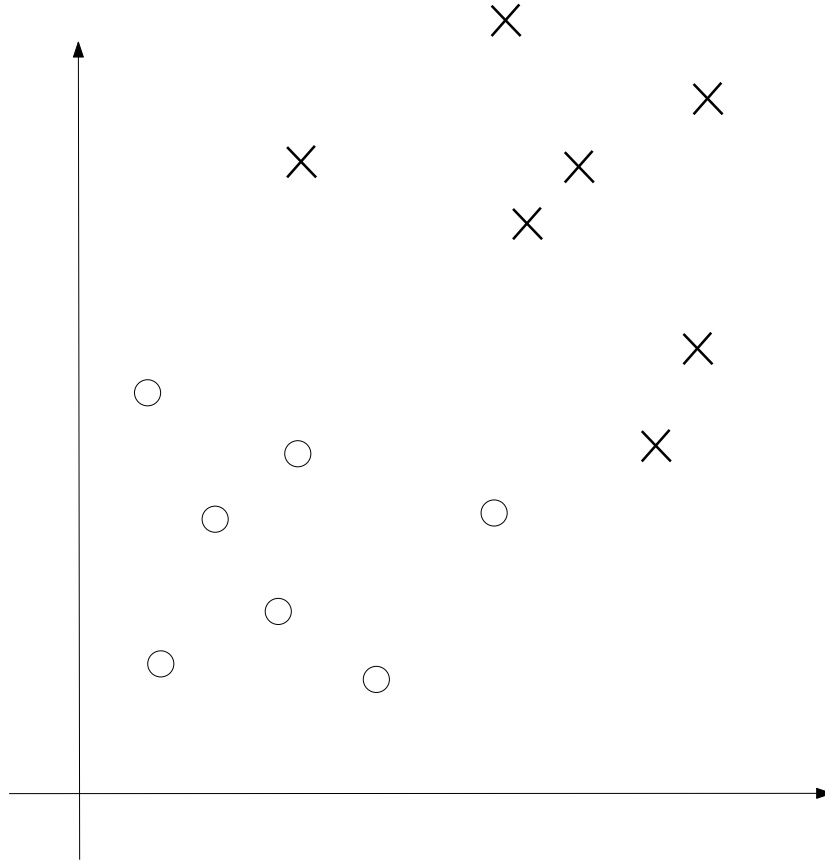
2. [2pts] If you are asked to use the Perceptron algorithm to classify these data points, what will the decision boundary look like? Explain the similarities and differences between Perceptron and SVM.

3. [2pts] Draw the SVM decision boundary for another set of data points on the following figure, and also explain the reason for achieving such a boundary. Please simply describe any other additional conditions or constraints you want to use, for example slack variable.
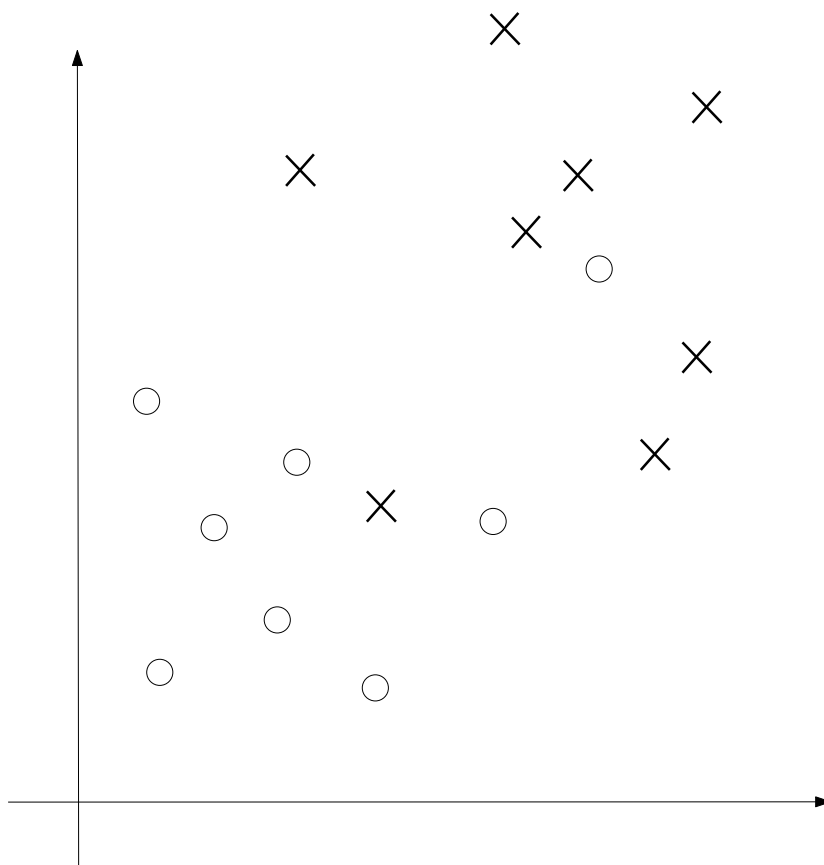
8 Compare Classifiers [6pts]

1. [2pts] When calculating confidence interval for a classifier, when we would like to use 10-fold cross-validation instead of holding out half of data?

2. [2pts] Given the sample mean and variance, calculating the p-value. Is the one-tailed p-value larger or the two-tailed p-value?

3. [2pts] Kevin proposed a hypothesis with accuracy 80% on a binary classification problem. While when you tested on 100 data points, you get accuracy for 70%. You can compute 95% confidence interval by

$$[Accu_S(h) - Z_N\sqrt{\frac{Accu_S(h)(1 - Accu_S(h))}{n}}, Accu_S(h) + Z_N\sqrt{\frac{Accu_S(h)(1 - Accu_S(h))}{n}}]$$

$$\text{where } Z_{0.95} = 1.96$$

Under 95% confidence level, do you believe Taiti's claim? Why?

9 Ensemble classifiers and semi-supervised learning [10 pts]

1. [2pts] The instances in Figure 7 and Figure 8 are actually user profiles on an on-line gaming site, where for a user, x_1 is the number of hours spent per week logged on, and x_2 is the number of unique games played. You would like to find out if you can predict, from this information, some demographical information, namely whether the user is male or female. About 5% of the users have created profiles including this information. What methods would be *plausible* to use in this case?

   (a) PCA

   (b) k-means

   (c) seeded k-means

   (d) seeded mixtures of Gaussians

   (e) seeded k-nearest-neighbor

2. [2pts] You suggest adding using additional features from a user's profile to make this task easier: specifically his/her age, how long he/she has had an account, and the specific games he/she have played. Kevin says that age is unnecessary to use as a feature, because you are trying to predict sex, and age and sex are independent. Is he right? Why or why not?

3. [2pts] Would it be more appropriate to use a transductive semi-supervised learner, or an inductive semi-supervised learner? Why?

4. [2pts] Which method is easier to parallelize: boosting, or bagging? Why?

5. [2pts] You have implemented stacking using 5 base classifiers and 10-fold cross-validation. One of the five base classifiers is used as the final meta-classifier (the "top" of the stack). If

each of your classifiers takes about the same time T to train on your dataset, and time t to evaluate a single example, how long does it to train the stacked learner? How long does it take to evaluate an example with the learned classifier? Assume nothing has been parallelized.
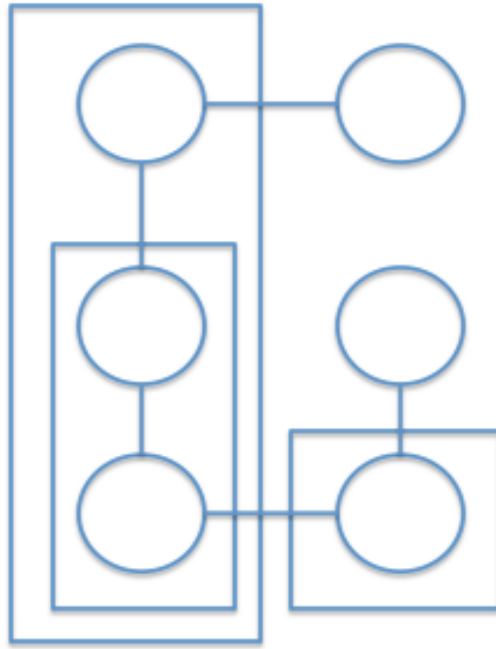
Figure 9: LDA plate diagram

10   LDA [12pts]

The figure 9 is an incomplete plate diagram for lda. We would like you to complete this. It should model a corpus with N documents each of length L, using K topics. Assume that α represent the prior for topic multinomial, θ_i the topic distribution for document indexed i, Z_{ij} is the topic indicator for word W_{ij}, β is the prior for word multinomial for each topic and ϕ_k is the word word multinomial for topic k.

1. [6pts] Draw the counts for each plate, the direction for each arrow and the variable for each circle.

2. [2pts] List the variables that are known in a typical use of lda

3. [2pts] List the variables that have a discrete domain.

4. [2pts] Name one inference method that can be used for lda.