Notes regarding my Questions:

1. 1p and 2p questions should be able to be answered regardlessly and may require writing just few sentences
2. 3p questions can be tough, maybe out of the scope of the exam (?) - maybe be a nice mental workout

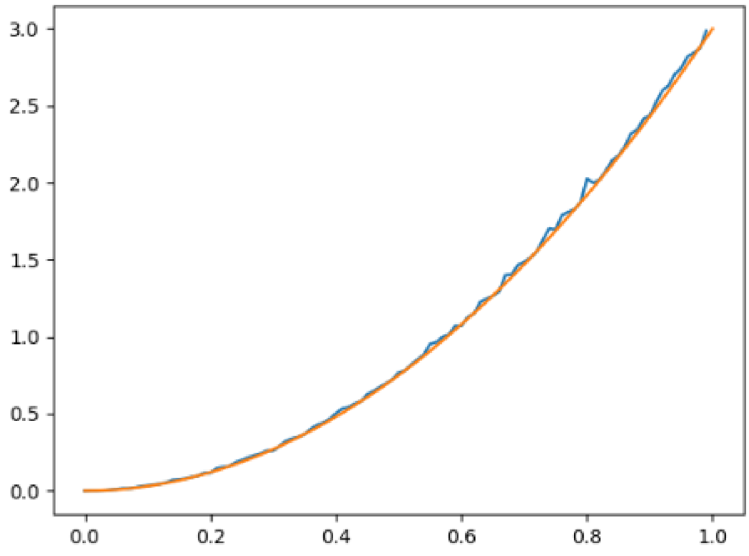
Questions:

1. Vocabulary

   Notes: Besides the inverse sampling theorem, there is not much happening in this chapter. One might want to look into multiple choice questions; for example several graphs are visualized (PDFs, CDFs, histograms etc.) and should now be matched together. This seems most likely.

   a. Formulate the sampling algorithm (2p)
   b. Compute the transform for a mapping from a uniform distribution to the target distribution 3*x^2, for the sake of simplicity we only care about the interval [0..1] (2p)
   c. Given me a uniform random distribution and a target distribution. Now we plug in **N** randomly chosen numbers from the uniform distribution into our mapping to the target distribution. We then collect the output and put it into a histogram. How would you expect the histogram to look like? You may assume **N** is big and the histogram has a suitable number of bins. (3p)

   E.g. our target PDF is 3*x^2 and the mapping is (3 * u) ** (1.0 / 3), view image below



2. Density Estimation
    a. Name a parametric and a non-parametric density estimation (1p)
    b. Why/how do we use density estimation with this method? (1p)
    c. What is a advantage of parametric over non parametric density estimation? (or vice versa) (1p)
    d. How can the class-conditional marginal densities for the Bayesian classifier be estimated? (2p)
    e. Are all kernels real valued and non-negative? (1p)
    f. When/why would you use histograms? (1p)
    g. How does the number of bins affect bias and variance? What is undersmoothing / oversmoothing? (1p)
    h. Name a few kernels. (2p)
    i. Which kernel minimalizes the MSE? (1p)
    j. How can the parameters of the parametric density function be estimated? Name two ways to do that (2p)
    k. What is the model selection problem? How can CrossVal and MaxLikeli be used? (2p)
3. Bias and Variance
    a. What is the bias-variance tradeoff? (1p)
    b. What is the bias error? What is the variance error? What is the generalization error? (2p)
    c. Do Bias and Variance require an error metric (like MSE)? (1p)
    d. Identify low/high bias/variance (2p).
    e. They are usually only looked at in the context of supervised learning, but not unsupervised learning. True/False? (1p)
    f. How could bias and variance nonetheless be used in the context of unsupervised learning? How could they be defined? (3p)
4. Random Forests and their Variants
    a. What is the idea of CARTs? (2p)
    b. Give the steps to train a Random Forest? (2p)
    c. Why do we prefer gini / cross entropy over missclassification rate (0p) ?
    d. How can we counter overfitting? (1p)
    e. Why do we use decision trees? What are their advantages? (2p)
    f. Write down 2 objective functions for classifications. (2p)
    g. Write down 1 objective function for regression. (1p)
    h. What is the gini index and what does it minimizes? (2p)
    i. What is the cross entropy and what does it minimizes? (2p)
    j. What is bagging? (1p)
    k. How does a RF get its smooth decision boundary? (1p)
    l. List 3 methods to decrease the variance in RF. (1p)
    m. How can we use RF for density estimation?
    n. Do we actually need bagging? Can we not train all decision trees of a random forests on the same dataset? (2p)
    o. What are density forests? (2p)

p. How do the number of the tress and the max depth of the trees influence bias and variance? (2p)
q. Can an increase of the depth of a decision tree also increase its training error? (1p)
r. What is the difference RF and Extreme Random Forests? (1p)
s. How can the features be ordered according to their importance? What is the out of bag error? (3p)

5. Introduction to Simplifications of the Feature Space
   a. What are the advantages of simpler feature spaces? Why would we want that? Name as many points as you want. (2p)
6. K-Means
   a. What is K-means? How does the algorithm work? (2p)
   b. What is the Within cluster distance? (1p)
   c. What does it mean, if we say: "K-means is locally optimal" (1p)
   d. What is Voronoi tesselation? (1p)
   e. Draw a pretty picture to illustrate how K-means can fail for a non-convex datasets and describe it. (2p)
   f. Something picture given. Perform one iteration of the K-Means clustering. (1p)
   g. Given some dataset (Swiss Roll), do K-means clustering or GMM seem sensible to be used? (1p)
   h. Does K-Means always converge? Explain your reasoning in detail. (3p)
7. Gaussian Mixture Models
   a. What is it? How does it work? (1p)
   b. Is EM guaranteed to converge? (1p)
   c. Does it converge to local or global optima? Does initialization matter? (1p)
   d. How do we achieve a good performance / result? (2p)
   e. Why do Gap Statistics not make sense for GMMs? (2p)
8. Mean Shift
   a. What is it? How does it work? (1p)
   b. What is the mean shift vector? (1p)
   c. What is the connection between the kernel size and the number of modes? (1p)
   d. Some picture are drawn: Which algorithm would you use here: "GMM, Mean shift or K-Means"
9. Model Selection for K-Means
   a. How can we find an optimal or sensible number of clusters for a given dataset? Name and explain such a method (2p)
   b. How does gap statistics work? (2p)
   c. What is a problem relying only on Gap statistics? (1p)
10. Model Selection for GMMs
    a. What is GMM? How does it work? (2p)
    b. How is overfitting avoided? (1p)

11. Curse of dimensionality
    a. Explain: What is the curse of Dimensionality (CuD), why is it a problem and how can we deal with it. (2 points)
    b. What are problems with dimensionality reduction? (2p)
    c. In a high dimensional space the data points likely become linearly separatable, thus high dimensionality is not necessarily a curse. True/False (2p)
    d. Given be a dataset that is completely contained in a d-dimensional hypercube of side length 1. Given now a d-dimensional unit sphere inside the cube, show that almost all data points, lie outside of the sphere. (Assuming uniform distribution of the data)(1p)
    (If you really feel like it, you can go and try computing the volume of the d-sphere. But there is a nicer way to go around it.)

12. Principal Component Analysis
    Notes: Question d) and e) check the fundamentals of Linear Algebra and if one knows what PCA actually does.
    a. Is PCA linear? (1p)
    b. What is its objective function? (1p)
    c. What is the connection between PCA and sum of mean squared errors? (2p)
    d. We now use PCA to throw away the two least important features, however in a slightly different fashion:
        i. We use PCA only once and throw away immediately the least important 2 features.
        ii. We use PCA twice. First we use it to remove the least important feature, we use PCA again and throw away again the current least important feature
    Do both verions result in the same output dataset? (2p)
13. Multidimensional Scaling
    a. Is MDS linear? (1p)
    b. Explain the steps in MDS.
    c. How are MDS and PCA similar? (1p)
14. ISOMAP
    a. What is Isomap? How does Isomap work? (2p)
    b. Is Isomap linear? (1p)
    c. What are geodesic distances? (1p)
    d. Can we use either floyd warshall or an iterated version of Dijsktra to compute all shortest pathes? (1p)
    e. Is isomap more robuts to noise than LE? (1p)
    f. Is isomap more robuts to outliers than LE? (1p)
15. Laplacian Eigenmaps
    a. What are LEs? How does they work? (2p)
    b. Are they linear? (1p)
    c. If we repeated the steps of 12d, would we get different results if used LEs instead of PCA? (2p)

16. Short Recap and Exam Remarks
17. Introduction to Probabilistic Graphical Models

    Notes: I haven't seen really much what can be asked here. One should look into the exercises and slides and do a few examples.
    If one is really into programming, then he/she may want to look into the RevBayes programming languages. Apparently it's made for working with graphic models.
18. Hidden Markov Models
    a. What is a Markov Process? (1p)
    b. What is a first order Markov Model? (1p)
    c. How is the number of parameters kept tractable? (1p)
    d. What is the Markov Assumption? (1p)
    e. What are some of the hyperparamters of HMMs? (1p)
19. HMM Algorithms 1 and 2
    a. What is the matching algorithm? (2p)
    b. What is the Viterbi algorithm? (2p)
    c. How do the forward/backward algorithm work? (2p)
20. HMM Algorithm 3
    a. How can HMM be trained via EM? (2p)
    b. Explain the methods of moments. (2p)
    c. Given be the following situation: You have a growing dataset size and tasked to train a HMM model on. Is EM a suited for consistendly training? (1p)
    d. What is an ergodic HMM? (1p)

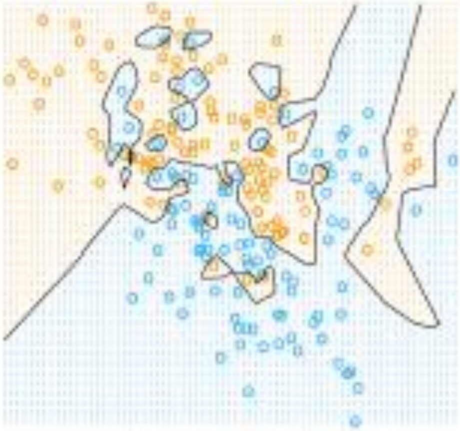
    General Note: EM belongs to the class of hill climbing algorithms, which in general suffer from local optima, ridges, alleys, zig-zagging, plateaus and in the case for EM also from inconsistency. Also view the joint sessions.

21. Conditional Independence
    a. One should have a look into the exercises and examples in the slides. I would expect something similar, nothing fancy.
    b. Something graph show something and something else is independent. (1p)
    c. Show something and something else is conditionally independent. (1p)
    d. Name all cliques in something graph. (1p)
22. Markov Random Fields

    Notes: Finding a good, suitable question here seems tricky.

23. Max Flow and Min Cut
    a. Simply have a look into the lecture slides, exercises and YT etc.
24. MRF Inference via Min Cuts
    a. What is the \alpha-expansion algorithm? (1p)
    b. What is the submodularity condition? (1p)
    c. How can Min Cuts be applied for MRF, e.g. denoising an Image? How does it work? (2p)
    d. Explain how to denoise an image using MRF , explain how to extend to gray scale images filtering (2p)

(e) [2 pts] Which of the these classifiers could have generated this decision boundary?



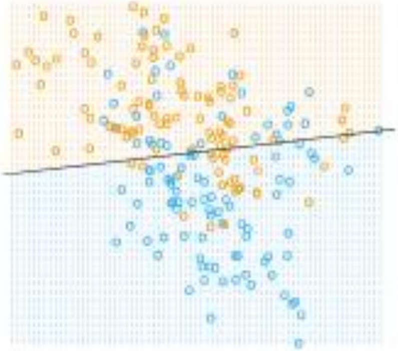
○ Linear SVM

○ Logistic regression

○ 1-NN

○ None of the above

(f) [2 pts] Which of the these classifiers could have generated this decision boundary?



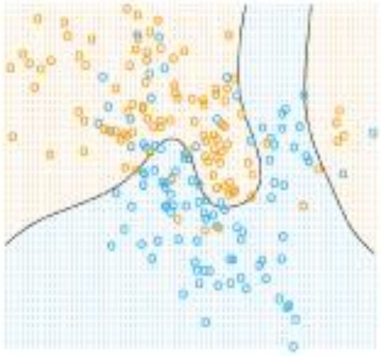
○ Linear SVM

○ 1-NN

○ None of the above

○ Logistic regression

(g) [2 pts] Which of the these classifiers could have generated this decision boundary?



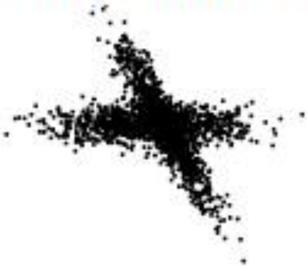
○ Linear SVM

○ 1-NN

○ None of the above

○ Logistic regression

(h) [2 pts] You want to cluster this data into 2 clusters. Which of the these algorithms would work well?

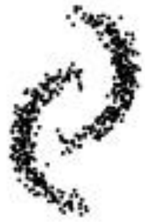


○ K-means

○ GMM clustering

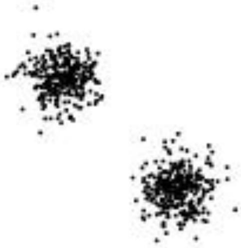
○ Mean shift clustering

(i) [2 pts] You want to cluster this data into 2 clusters. Which of the these algorithms would work well?



○ K-means ○ GMM clustering ○ Mean shift clustering

(j) [2 pts] You want to cluster this data into 2 clusters. Which of the these algorithms would work well?



○ K-means ○ GMM clustering ○ Mean shift clustering