Academic year 2017-2018, January 22nd, 8h30

Machine learning

Clearly write your name on every answer sheet, as well as on the question sheet. Answer each question separately, and clearly mention the number of every question next to your answer. If you don't provide an answer for a question, clearly mention the question number and write "No answer".
No laptops, calculators, PDAs, phones or Internet access is allowed. Hand in all question sheets.

---

1. **Bias-Variance tradeoff (6 pt)**

   (a) Figure 1 describes a general picture of the bias-variance tradeoff of a classifier.

   - Which of the curves is more likely to be the training error and which is more likely to be the validation error? Indicate on the graph by filling the dotted lines.
   - In which regions of the graph are bias and variance low and high? Indicate clearly on the graph with four labels: "low variance", "high variance", "low bias", "high bias".
   - In which regions does the model overfit or underfit? Indicate clearly on the graph by labeling "overfit" and "underfit".

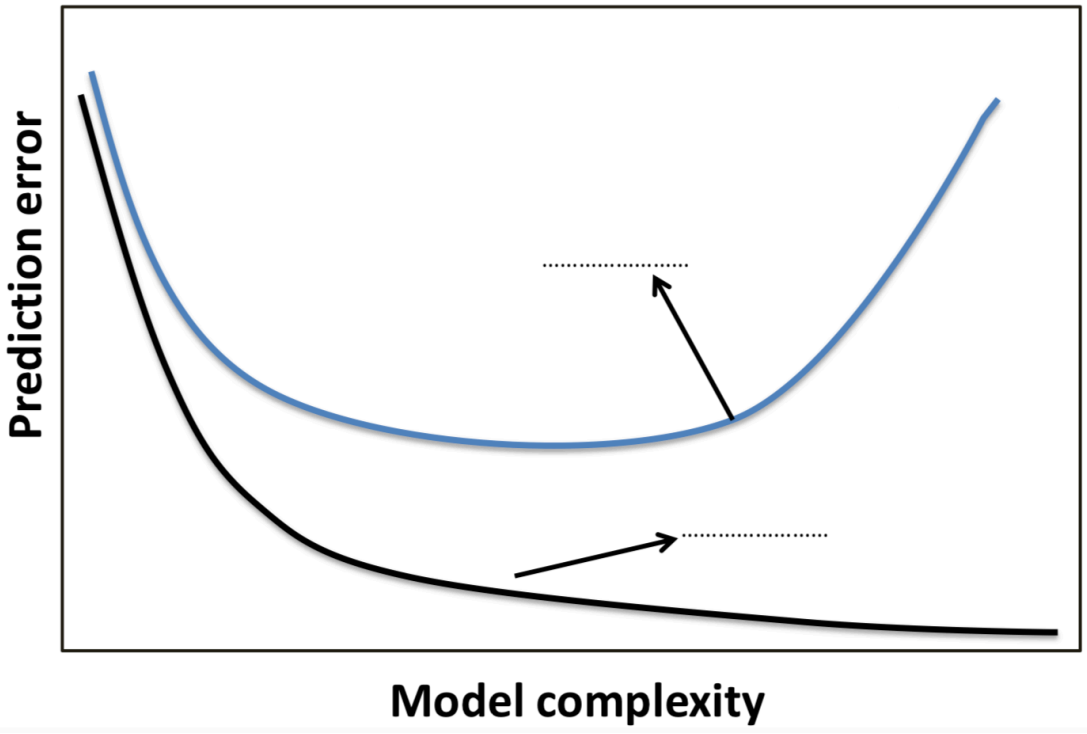


Figure 1: Bias-variance tradeoff

(b) A set of data points is generated by the following process: $Y = w_0 + w_1 X + w_2 X^2 + w_3 X^3 + w_4 X^4 + \epsilon$, where X is a real-valued random variable and ϵ is a Gaussian noise variable.
You use two models to fit the data:
Model 1: $Y = aX + b + \epsilon$
Model 2: $Y = w_0 + w_1 X^1 + \ldots + w_9 X^9 + \epsilon$

Which of the following is true:

i. Model 1, when compared to Model 2 using a fixed number of training examples, has a bias which is:
   A. Lower
   B. Higher
   C. The same

ii. Model 1, when compared to Model 2 using a fixed number of training examples, has a variance which is:
   A. Lower
   B. Higher
   C. The same

iii. Given 10 training examples, which model is more likely to overfit the data ?
   A. Model 1
   B. Model 2

(c) Explain the effect (increase, decrease, no change) of the following actions on the bias and variance:

i. Reducing the number of leaves in a decision tree

ii. Increase K in a K-Nearest Neighbor classifier

iii. Increase the number of training examples in linear regression

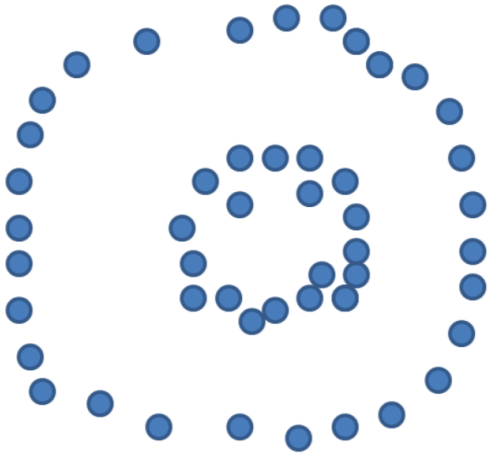
2. **Clustering (6 pt)**

(a) Compare the advantages and disadvantages of DBScan to Kmeans

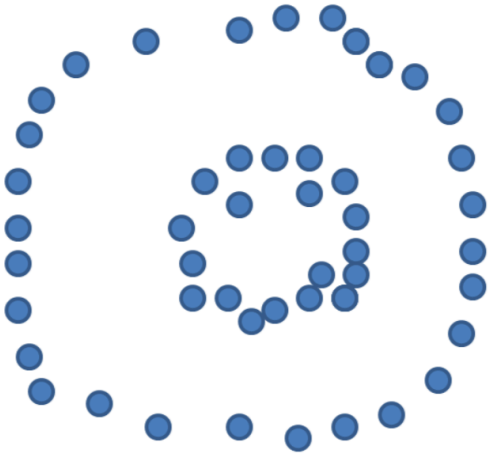
(b) Imagine we would like to cluster houses around Ghent without using their exact addresses. For each house, we map properties of the house to a numeric value. For instance, the house's location is mapped as Center = 0, Ledeberg = 1, Gentbrugge = 2, etc., the exterior material is brick = 0, aluminum = 1, wood = 2, etc., the kitchen color is white = 0, green = 1, tan = 2, etc. We have 50 such features so each house can be represented as a vector in $\mathbb{R}^{50}$. Which of the following three clustering algorithms (hierarchical clustering, k-means and Gaussian mixture models) would be most appropriate for this task? Explain briefly for each algorithm.

(c) We would like to cluster the points in Figure 2a and Figure 2b (which are the same) using k-means and GMM, respectively. In both cases we set k = 2. We perform several random restarts for each algorithm and choose the best one. For each method show the resulting cluster centers in the appropriate figure (k-means on Figure 2a and GMM on Figure 2b).

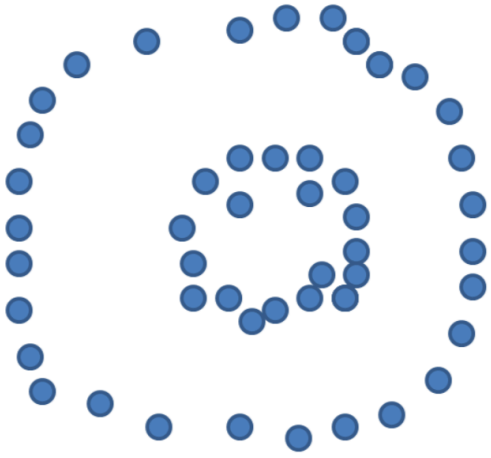
(d) For the same figure (which is repeated in Figures 2c and 2d) we would like to use hierarchical clustering. We will use the Euclidian distance as the distance function. In both cases we cut the tree at the second level to obtain two clusters. For two of the linkage models learned in class, single and average link, circle the resulting groups of points on each of the figures (Figure 2c - single link, Figure 2d - average link).



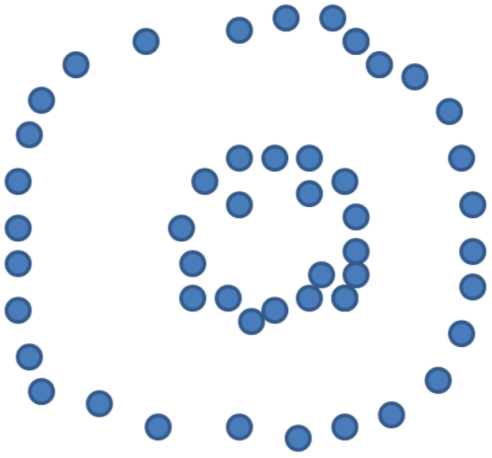
(a) KNN



(b) GMM



(c) Single link

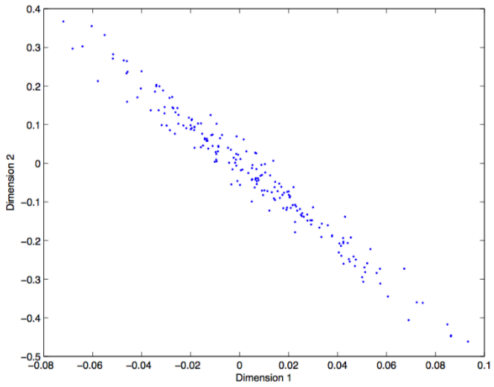


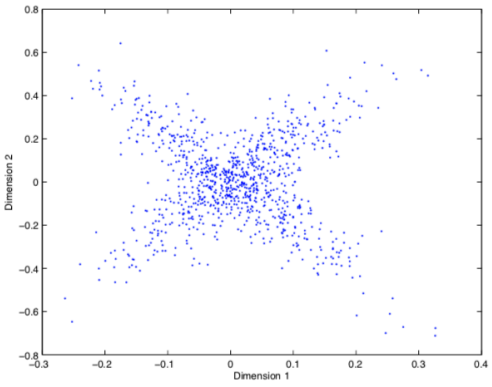
(d) Average link

Figure 2

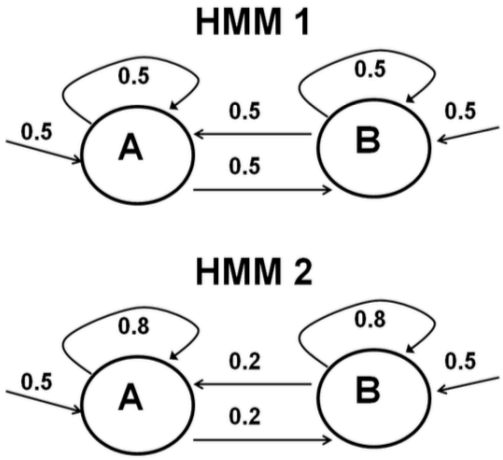
3. **Dimensionality reduction - PCA (1 pt)**

   Principal component analysis is a dimensionality reduction method that projects a dataset into its most variable components. You are given the following 2D datasets, draw the first and second principle components on each plot.





4. **Hidden Markov Models (2 pt)** The figure below represents two HMMs. States are represented by circles and transitions by edges. In both, emissions are deterministic and listed inside the states. Transition probabilities and starting probabilities are listed next to the relevant edges. For example, in HMM 1 we have a probability of 0.5 to start with the state that emits A and a probability of 0.5 to transition to the state that emits B if we are now in the state that emits A. In the questions below, $O_{100} = A$ means that the 100th symbol emitted by the HMM is A.



   (a) What is $P(O_{100} = A, O_{101} = A, O_{102} = A)$ for HMM1 ?

   (b) What is $P(O_{100} = A, O_{101} = A, O_{102} = A)$ for HMM2 ?

   (c) Let P_1 be: $P_1 = P(O_{100} = A, O_{101} = B, O_{102} = A, O_{103} = B)$ for HMM1 and let $P2$ be: $P_2 = P(O_{100} = A, O_{101} = B, O_{102} = A, O_{103} = B)$ for HMM2. Choose the correct answer from the choices below and briefly explain.

i. $P_1 > P_2$
ii. $P_2 > P_1$
iii. $P_1 = P_2$
iv. Impossible to tell the relationship between the two probabilities

5. **Linear Regression (2 pt)**
We are given a set of two-dimensional inputs and their corresponding output pair: $\{x_{i,1}, x_{i,2}, y_i\}$. We would like to use the following regression model to predict y:

$$y_i = w_1^2 x_{i,1} + w_2^2 x_{i,2}$$

Derive the optimal value for w_1 when using least squares as the target minimization function (w_2 may appear in your resulting equation). Note that there may be more than one possible value for w_1.

6. **True or False (3 pt)**
Are the following statements True or False ? If True, explain in at most two sentences. If False, explain why or give a counterexample in at most two sentences.

(a) Overfitting is more likely when the hypothesis space is small

(b) For data D and hypothesis H it is always true that $\sum_h P(H = h|D = d) = 1$

(c) For data D and hypothesis H it is always true that $\sum_h P(D = d|H = h) = 1$

(d) The following is a good procedure for performing feature selection. A project team performed a feature selection procedure on the full data and reduced their large feature set to a smaller set. Then they split the data into test and training portions. They built their model on training data using several different model settings, and report the best test error they achieved.

(e) Suppose X_i are categorical input attributes and Y is a categorical output attribute. Assume we plan to learn a decision tree without pruning, using the standard algorithm. If $IG(Y|X_i) = 0$ according to the values of entropy and conditional entropy computed from the data, then X_i will not appear in the decision tree.

(f) If X and Y are independent and $X > 1$, then $\text{Var}[X + 2Y^2]=\text{Var}[X]+4\text{Var}[Y^2]$ and $\text{E}[X^2 - X] \geq \text{Var}[X]$