Machine Learning I

Retake Exam

June 5th, 2020

1. Linear Classification

1A. Linear Models (3 points)

- [1p] What does it mean for a classification model to be is linear?
- [1p] Give 3 examples of linear classification models
- [1p] How are they different from each other?

1B. Linear Discriminant Analysis (5 points)

Given the dataset from table 1

- [3p] Fit a LDA model using the labeled datapoints, report the values of π_k, μ_k, Σ_k and $\hat{\Sigma}$.
- [2p] Predict the class of the unlabeled datapoint using the trained model.

x	y	class
-1	1	A
-1	-1	A
0	1	A
0	-1	A
0	0	B
2	0	B
1	1	B
0	0	-

Table 1

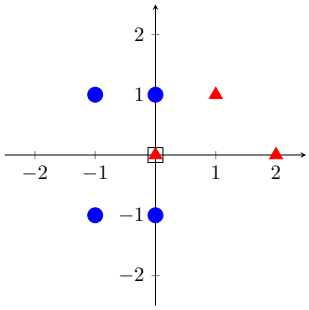
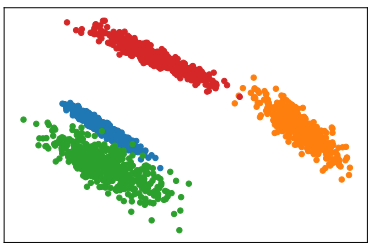


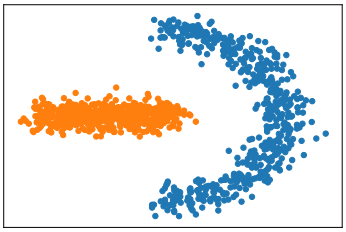
Figure 1

1C. LDA (2 points)

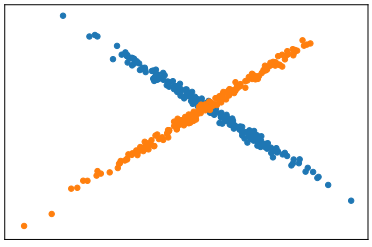
- [1p] Explain the difference between LDA and QDA.
- [1p] For each of the datasets depicted in Figures 2a,2b and 2c, decide which model is more appropriate: LDA or QDA. **Note:** You don't need to explain your decision



(2a)



(2b)



(2c)

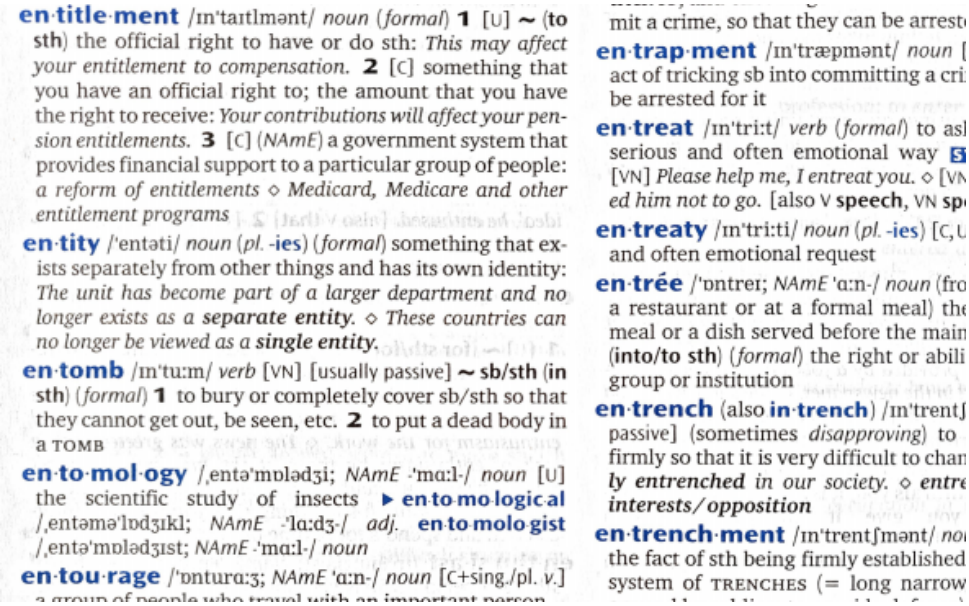
2. K-Nearest Neighbors

2A. KNN basics (3 points)

Name 4 different data types and what corresponding distance measures.

2B. KNN bases auto-correct (5 points)

You want to write an auto-correction software that corrects spelling mistakes when users write a text message. Your idea is to use a KNN model that provides the user with a list of the 1-nearest-neighbors using the LEVENSTHEIN distance.



en·title·ment /ɪn'taɪtlmənt/ *noun* (*formal*) **1** [U] ~ (**to sth**) the official right to have or do sth: *This may affect your entitlement to compensation.* **2** [C] something that you have an official right to; the amount that you have the right to receive: *Your contributions will affect your pension entitlements.* **3** [C] (*NAmE*) a government system that provides financial support to a particular group of people: *a reform of entitlements* ◇ *Medicard, Medicare and other entitlement programs*

en·tity /'entəti/ *noun* (*pl.* **-ies**) (*formal*) something that exists separately from other things and has its own identity: *The unit has become part of a larger department and no longer exists as a* ***separate entity.*** ◇ *These countries can no longer be viewed as a* ***single entity.***

en·tomb /ɪn'tu:m/ *verb* [VN] [usually passive] ~ **sb/sth** (**in sth**) (*formal*) **1** to bury or completely cover sb/sth so that they cannot get out, be seen, etc. **2** to put a dead body in a TOMB

en·to·mol·ogy /ˌentə'mɒlədʒi; *NAmE* -'mɑ:l-/ *noun* [U] the scientific study of insects ▸ **en·to·mo·logic·al** /ˌentəmə'lɒdʒɪkl; *NAmE* -'lɑ:dʒ-/ *adj.* **en·to·molo·gist** /ˌentə'mɒlədʒɪst; *NAmE* -'mɑ:l-/ *noun*

en·tou·rage /'ɒnturɑ:ʒ; *NAmE* 'ɑ:n-/ *noun* [C+sing./pl. *v.*]

mit a crime, so that they can be arrest

en·trap·ment /ɪn'træpmənt/ *noun* [act of tricking sb into committing a cri be arrested for it

en·treat /ɪn'tri:t/ *verb* (*formal*) to as serious and often emotional way [VN] *Please help me, I entreat you.* ◇ [VN *ed him not to go.* [also V **speech**, VN **sp**

en·treaty /ɪn'tri:ti/ *noun* (*pl.* **-ies**) [C, U and often emotional request

en·trée /'ɒntreɪ; *NAmE* 'ɑ:n-/ *noun* (fro a restaurant or at a formal meal) the meal or a dish served before the main

(into/to sth) (*formal*) the right or abili group or institution

en·trench (also **in·trench**) /ɪn'trentʃ passive] (sometimes *disapproving*) to firmly so that it is very difficult to chan ***ly entrenched*** *in our society.* ◇ ***entre interests/opposition***

en·trench·ment /ɪn'trentʃmənt/ *no* the fact of sth being firmly established system of TRENCHES (= long narrow

Figure 3: dictionary excerpt

- [1p] Given the dictionary entries from Figure 3, identify all nearest neighbors to the word `entry`

  Note: You do not need to compute the Levensthein distance explicitly.

Part 2. A common mistake while typing with two hands is that the order of two letters gets swapped.

- [2p] Explicitly compute the Levensthein distance between `swap` and `swpa`.

Part 3. Consider the following example text

> Yesterday, I ate a delicious lie.

Here, the 1-nearest neighbour search in the dictionary yields the following list of matches

```
die, fie, hie, lie, pie, tie, vie, die, lee, lye, lib, lid, lii, Lin, lip, Lir,
lit, liv, Liz, lice, life, like, lime, line, lire, lite, live
```

However, we want to only propose 3 auto-correct recommendations to the user. To do this, we want to build a model that takes the nearest neighbor matches and returns a likelihood for each of them.

- [2p] Describe which additional features could be useful for such a model. Guess what the top 3 matches of a model trained with these features would be, and give a short argument why.

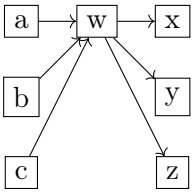
2C. Large Scale KNN (2 points)

In problem 2B the performance of the model can be slow if the dictionary is very large. How could the performance of the Nearest Neighbor model be improved?

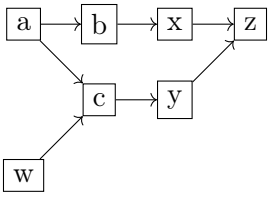
3. Bayesian Networks

3A. Bayesian Networks Basics (3 points)

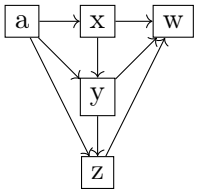
Provide the factorization of the joint distribution corresponding to the Bayesian Networks depicted in Figure 4a, 4b and 4c.



(4a)



(4b)



(4c)

3B. Naive Bayes Application (5 points)

customer	toilet paper	N95 mask	pasta	hand sanitizer
1	no	no	yes	no
2	no	no	no	yes
3	yes	yes	no	yes
4	yes	yes	yes	yes
5	yes	yes	yes	no
6	yes	no	yes	yes
7	-	yes	yes	yes

Table 2: Shopper preferences during COVID-19 pandemic

- [4pt] Train a Naive Bayes Classifier on the labeled datapoints from Table 2 for the target variable "toilet paper". Use add-one-smoothing for the Dirichlet prior. Provide the conditional probability tables for each node.
- [1pt] What is the probability that the last customer will buy toilet paper?

3C. Naive Bayes Theory (2 points)

Suppose that we had a dataset for which it is known that all the predictor variables $X_1, X_2, \ldots$ are categorical, **independent** random variables. In this case is it possible for any other classifier (e.g., a decision tree or a neural net) to do better than a naive Bayes classifier?

4. Clustering

4A. Unsupervised Learning (3 points)

- [2p] Briefly explain the difference between supervised and unsupervised learning
- [1p] Briefly explain the difference between hard and soft clustering

4B. K-medoids (5 points)

- [2p] Find the distance matrix of the dataset from table 5a / figure 5b w.r.t. the L^1 metric.

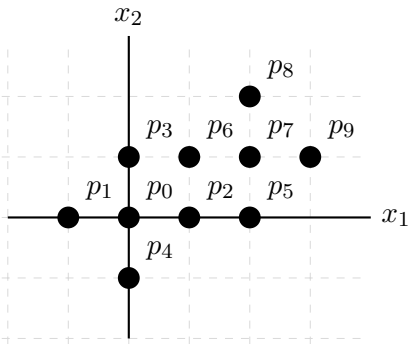
  Note: You do not need to provide explicit calculations for the individual distances.

- [3p] Apply K-medoids clustering to the dataset from table 5a / figure 5b, using $K = 2$ and the Manhattan/L^1 distance measure. Start with $\mu_A = p_1$ and $\mu_B = p_6$.

  Note: Perform up to 3 iterations, if the algorithm does not terminate earlier.

ID	x_1	x_2
p_0	0	0
p_1	−1	0
p_2	1	0
p_3	0	1
p_4	0	−1
p_5	2	0
p_6	1	1
p_7	2	1
p_8	2	2
p_9	3	1

(5a)



(5b)

4C. Clustering chess positions (2 points)

A company creating chess software has a large database of chess games. They want you to identify positions that occur in many games and group the positions accordingly. Come up with a distance measure that allows you to apply clustering to chess positions.

Note: If you do not know what chess is, you can use any other turn based board game (such as tic-tac-toe, checkers, monopoly, etc.). You **do not** need to explicitly verify the distance measure axioms.