Introduction to Information Theory and Machine Learning — Aug 27th, 2020.
(Prof. Permuter Haim, Mr. Eli Shmuel, Mr. Dor Tsur and Mr. Ben Marinberg)

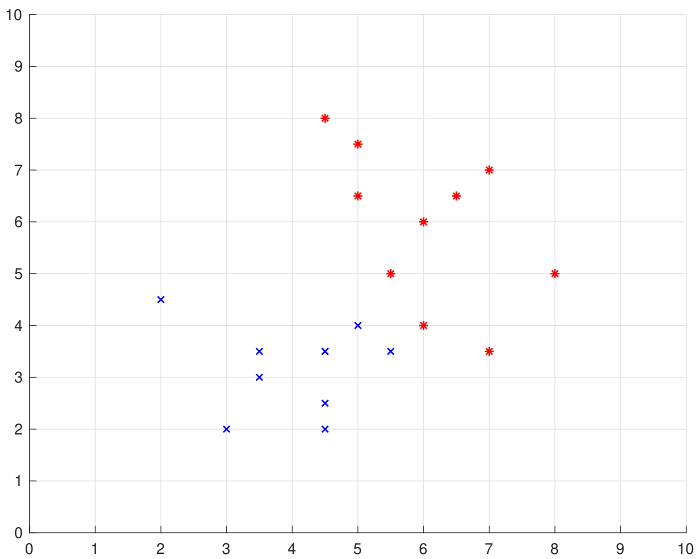
Final Exam - Moed Bet
Total time for the exam: 3 hours!

Important: Please copy the following sentence and sign it: " I am respecting the rules of the exam: Signature:__________ "

1) **Supervised and unsupervised learning (36 pt)**
You are given a training set of 19 samples in Fig 1. 10 red samples marked as red asterisks (*) and the rest are marked as blue Xs i.e., X. Each sample is given by a 2D feature vector. Create a unique test set by using the first 8 digits of your ID number, i.e. for ID number 123456789 the test set is$[(1,2),(3,4),(5,6),(7,8)]$

Feature 1	Feature 2	Label
2	4.5	Blue
3	2	
3.5	3	
3.5	3.5	
4.5	2	
4.5	2.5	
4.5	3.5	
5	4	
5.5	3.5	
4.5	8	Red
5	6.5	
5	7.5	
5.5	5	
6	4	
6	6	
6.5	7.5	
7	3.5	
7	7	
8	5	

Table form



2D Map form

Fig. 1: Training set

Supervised

a) Use the training dataset to classify your test set using KNN for $K = 1, 3$.
b) Fit a single Gaussian to each class
   i) Assume full covariance matrix and draw Gaussians' contour lines and the decision line.
   ii) Assume diagonal covariance matrices and draw the Gaussians' contour lines and decision line.

Unsupervised

Use $(4, 4)$ and $(7, 4)$ as the initial points for the following.

c) Fit a unsupervised GMM, assume diagonal covariance matrices. Draw the Gaussians' contour lines and decision line.
d) Fit a K-means to the data and draw the outcome centroids and decision line.

Ranking the models

e) Rank the supervised models according to the complexity of the inference. Explain your answer.
f) Rank the unsupervised models according to the complexity of the inference. Explain your answer.

Solution:

a) Your may verify your answers for your unique test set using the region-divided figures on Fig2.
b) See Fig 3.
c) See Fig4.
d) See Fig 5
e) Inference complexity from small to large, Assuming the data set is sufficiently large:
   GMM-diagonal $\rightarrow$ GMM-full $\rightarrow$ 1NN $\rightarrow$ 3NN.
   This is because per sample GMM need to calculate the probability for each class while KNN need to run over all samples.
f) Inference complexity: Kmeans $\rightarrow$ GMM-diagonal.
   Kmeans needs to find the closest centroid while GMM calculates probability (exponentiation, matrix inverse, etc.). When the covariance matrix is αI_n, where α is constant and I_n is identity matrix of size n, then the complexity of Kmeans and GMM is the same (and will yield similar predictions).

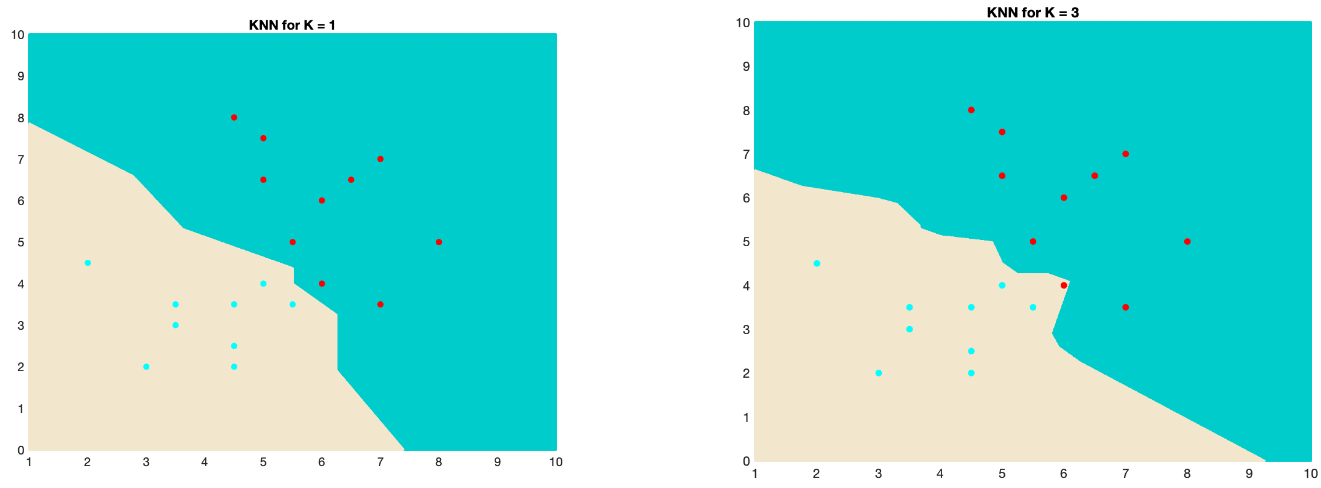


Fig. 2: KNN classification regions for K=1,3

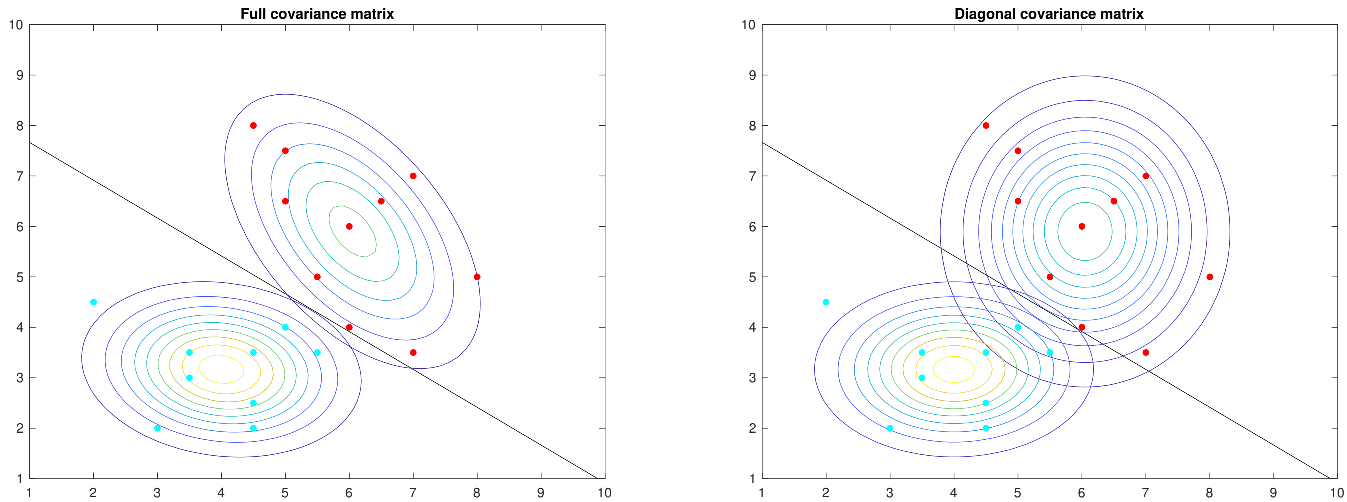


Fig. 3: Supervised GMM using diagonal/full covariance matrices

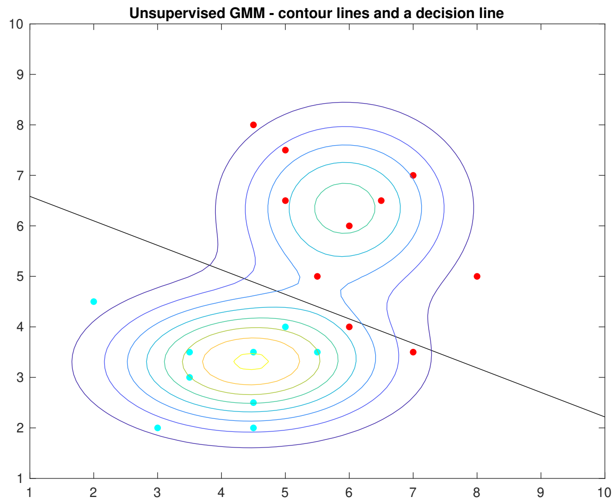


Fig. 4: GMM with 2 classes and diagonal matrices

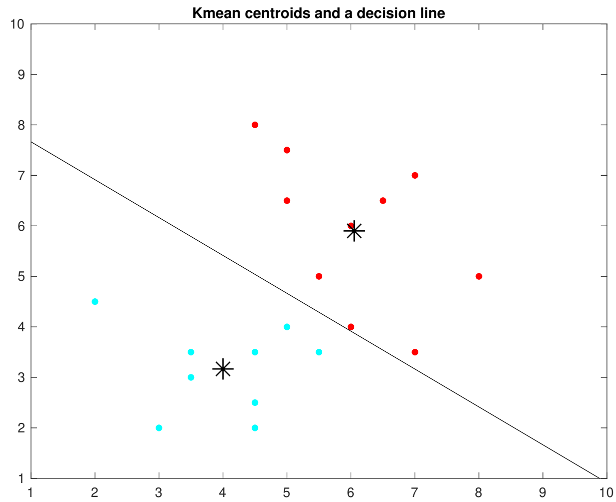


Fig. 5: K-means

Fig. 4-5: Unsupervised models initiated from (4,4) and (4,7)

2) **MINE (23 points)**
   a) **(10 points)** Suppose we are given with a set of i.i.d. samples from two random variables, $\{(x_i, y_i)\}_{i=1}^n$, where $(x_i, y_i,) \sim P_{X,Y}, \quad x_i, y_i \in \mathbb{R}$ and we attempt to estimate $I(X;Y)$. The method is based on MINE $I_n(X,Y)$ as taught in class.

Which of the following experiment graphs is a possible description of the estimated MI during training (Figure 6)? Write the indices of the correct graphs in your solution and explain your choice.

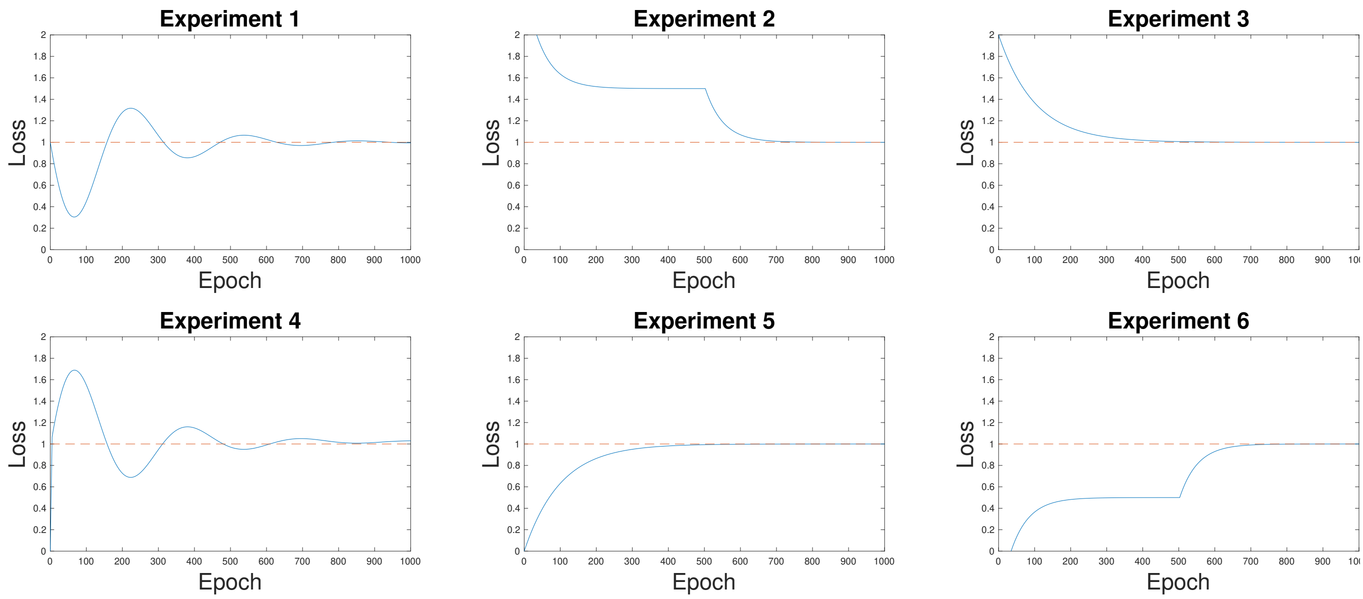


Fig. 6: Experiments - filled line represents the model's loss and dashed line represents the true value.

Solution: As learned in class, the objective of MINE is a lower bound on the mutual information, due to the supremum. Therefore, only graphs 5 and 6 are valid. The answer of choosing graphs 3 and 2 instead was also accepted, but note that this answer is not completely valid, because implementation of of this optimization with a minus sign might lead to descending cost, but then we would expect the limit to be $-I(X;Y)$.

b) Now, we wish to estimate the multivariate mutual information between the d-dimensional random vectors, (X^d, Y^d), based on the algorithm and 1-dimensional samples presented in (2a), .
   i) **(8 points)** What criteria should the elements of $X^d, Y^d \in \mathbb{R}^d$ satisfy if we want to use the dateset given in question (2a)?
   ii) **(5 points)** What is the relation between the output of MINE and $I(X^d;Y^d)$ under this criteria and $n \to \infty$?

Solution We need to find a way to express $I(X^d, Y^d)$ with $I(X;Y)$, the limit of a trained MINE model with the sample set of subsection (a). We can use the chain rule and result with

$$\begin{aligned} I(X^d;Y^d) &= \sum_{j=1}^{d} I(X^d;Y_j|Y^{j-1}) \\ &= \sum_{j=1}^{d}\sum_{k=1}^{d} I(X_k;Y_j|X^{k-1}Y^{j-1}). \end{aligned}$$

Therefore, we need to demand

$$I(X_k;Y_j|X^{k-1}Y^{j-1}) = I(X_1;Y_1) \quad j,k = 1,\ldots,d. \tag{1}$$

Which happens iff the components of (X^d, Y^d) are iid. Under this criteria we could use MINE $I_n(X,Y)$, the estimator of $I(X;Y)$ to calculate the d-dimensional mutual information, and the relation is given by:

$$I(X^d;Y^d) = dI_n(X,Y) \tag{2}$$

3) **Bound on the binomial coefficient and types. (51 points)**
Let $k, n \in \mathbb{N}$ such that $k \leq n$. Let $q := \frac{k}{n}$. **Reminder:** The binomial expression is given by

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} \tag{3}$$

Furthermore the number of sequences of length n with k ones is $\binom{n}{k}$.
In this question we prove the following inequality:

$$\frac{1}{n+1} 2^{nH_b(q)} \leq \binom{n}{k} \leq 2^{nH_b(q)} \tag{4}$$

where, $H_b(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy function.
Steps:

a) **(3 points)** Show that:

$$1 = \sum_{i=0}^{n} \binom{n}{i} q^i (1-q)^{n-i} \tag{5}$$

Solution:
From the binomial expression:

$$\sum_{i=0}^{n} \binom{n}{i} q^i (1-q)^{n-i} = (q + (1-q))^n = 1 \tag{6}$$

b) **(5 points)** Deduce that for every $i = 0, \ldots, n$:

$$\binom{n}{i} \leq \frac{1}{q^i (1-q)^{n-i}} \tag{7}$$

Solution:
We note than every argument of the sum $\sum_{i=0}^{n} \binom{n}{i} q^i (1-q)^{n-i}$ is **non-negative** (since for every i: $\binom{n}{i}, q, (1-q) \geq 0$), and according to b) the sum is equal to 1. Hence, any argument by itself $\binom{n}{i} q^i (1-q)^{n-i} \leq 1$.

c) **(5 points)** Using the above inequality show that the right hand side of inequality (4) holds.
Solution:
After simplification using log rules and $q := \frac{k}{n}$ we have:

$$2^{nH_b(q)} = \frac{1}{q^k (1-q)^{n-k}}. \tag{8}$$

Hence, the right hand side of inequality (4) holds according to (7) because $k \in \{0, 1, ..., n\}$.

d) **(8 points)** Assume the following:

$$\text{argmax}_{i=0,\ldots,n} \binom{n}{i} p^i (1-p)^{n-i} = np, \quad np \in \mathbb{Z} \tag{9}$$

prove the following inequality:

$$\binom{n}{k} \geq \frac{1}{n+1} 2^{nH_b(q)} \tag{10}$$

Solution:
In our case, $\text{argmax}_{i=0,\ldots,n} \binom{n}{i} q^i (1-q)^{n-i} = nq = k$, hence

$$1 = \sum_{i=0}^{n} \binom{n}{i} q^i (1-q)^{n-i} \leq \sum_{i=0}^{n} \binom{n}{k} q^k (1-q)^{n-k} = (n+1)\binom{n}{k} q^k (1-q)^{n-k} = (n+1)\binom{n}{k} 2^{-nH_b(q)}, \tag{11}$$

where the last equality follows from (8). From the positivity of the arguments we can divide and get (10).

e) **(8 points)** The type of a Binary sequence of length n with k ones is $\frac{k}{n}$. Provide a lower and upper bound on the number of sequences of length n of a specific type $\frac{k}{n}$.
Solution:
As it can be deduced from c) and d), (4) holds.

f) **(3 points)** Is the number of sequences with a fixed type exponential or polynomial?
Solution:
Exponential. We test the lower and upper bounds of (4) when $n \to \infty$ while q is fixed (consequently $H_b(q)$ is fixed). $2^{nH_b(q)}$ is exponential, while for the lower bound the denominator $n+1$ is only polynomial, hence both bounds are exponential, then $\binom{n}{k}$ is exponential.

g) **(8 points)** How many different types there exists if the sequence is Binary and has length n.
Solution:
Summing all possible types $\frac{k}{n}$ of length n with $k \in \{0, 1, ..., n\}$ ones, we have $n+1$ different types.

h) **(3 points)** Is the number of different types that you calculated in the previous sub-question exponential or polynomial?
Solution:
Polynomial.

i) **(8 points)** For a binary sequence of length n with a type p we actually consider the ratio $\frac{k}{n}$ of such that $\frac{k}{n}$ is the closest number to p. For instance if $p = 0.5$ and $n = 100$ or $n = 101$ then $k = 50$. How many sequences of a fixed type p there exists asymptotically, namely, calculate

$$\lim_{n \to \infty} \frac{1}{n} \log(\text{number of sequences of type } p).$$

Solution:

We examine this limit for both the lower and upper bounds. We note that $\lim_{n\to\infty} q(n,p) = \lim_{n\to\infty} \frac{k(n,p)}{n} = p$. Then,

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \log(\frac{1}{n+1} 2^{nH_b(q(n,p))}) &= \lim_{n\to\infty} \frac{1}{n} \log\left(\frac{1}{n+1}\right) + \frac{1}{n} \log(2^{nH_b(q(n,p))}) \\ &= \lim_{n\to\infty} \frac{1}{n} \log(2^{nH_b(q(n,p))}) \\ &= \lim_{n\to\infty} H_b(q(n,p)) \\ &= H_b(p), \end{aligned} \tag{12}$$

and this is the limit for **both** bounds. Finally, from the sandwich theorem we have that the limit in the question equals $H_b(p)$.

Intuition: In this question you have developed an analysis of the number of sequences with a fixed ratio of the alphabet. Note, that you have obtained the entropy expression as a result of a basic combinatorical question. This result is very fundamental and requires only basic knowledge. Hence, we could actually start the course and define the entropy using this fundamental result.

Good Luck!