CSE 512: Final review

---

1. **Additional nomenclature** Define the following terms, and describe differences between each other
   (a) Discriminative vs generative models
   (b) Statistical inference
   (c) Ensemble learning, bagging, and boosting
   (d) one-vs-one, one-vs-all for multiclass classification
   (e) clustering vs classification
   (f) DAG, HMM, graphical models
   (g) representation learning, why/how we use it
   (h) dimensionality reduction, PCA, ICA
   (i) Strong duality, weak duality, Lagrangian saddle point problem, Lagrange dual variables
   (j) entropy and information gain
   (k) KNN, Kmeans, and GMM
   (l) Kernal SVM vs linear SVM vs dual SVM
2. **Understanding decision trees**
   (a) Using the training data below, construct a decision tree by following the steps below.

animal	label	location	color	outer
Hello	cat	house	white	fur
Garfield	cat	house	orange	fur
Tom	cat	house	blue	fur
Butch	cat	alley	black	fur
Babe	pig	farm	pink	skin
Wilbur	pig	farm	pink	skin
Nemo	fish	ocean	orange	scales
Ursula	fish	ocean	black	skin
Marlin	fish	ocean	orange	scales
Mighty	armidillo	jungle	red	shell

      i. What is the entropy of the labels?
      ii. For the first split, pick the feature that maximizes information gain. Each split should only have 2 children for yes or no, e.g. house? yes or no is different than ocean? yes or no, even though they have the same feature.
      iii. Which node has the least purity?
      iv. Split on the least purity node again, and continue until the decision tree is made. At each step, report the information gain.
   (b) Using this tree, infer the labels of the animals in the following test set and report the test error.

animal	true label	location	color	outer
Pumbaa	pig	jungle	pink	skin
Ozone	cat	alley	pink	skin
Salem	cat	house	black	fur

   (c) **Conceptual** Name 3 hints that a decision tree for classification has overfit
3. **Understanding regression trees**
   (a) Using the training data below, construct a decision tree.

person	height (in)	favorite sport	age	consumption of milk
Alice	61	swimming	24	frequent
Brian	52	swimming	10	rare
Carlos	73	basketball	18	average
Dianne	24	none	1.5	frequent
Esther	75	basketball	35	frequent
Freddy	36	basketball	7	frequent
Gary	59	none	15	average
Harris	65	swimming	48	average
Ivy	64	swimming	36	rare

i. What is the mean squared error of picking the average height as the prediction for all the datapoints?

ii. For the first split, pick the feature that maximizes information gain. Each split should only have 2 children; for the sport and consumption of milk, it should be a yes or no question. For age, it should be a value to threshold.

iii. Which node has the highest mean squared error?

iv. Split on this node again, and continue until the decision tree is made. At each step, report the new mean squared error.

(b) Using this tree, infer the labels of the animals in the following test set, and return the test mean squared error.

person	height (in)	favorite sport	age	consumption of milk
Joe	57	basketball	12	frequent
Katherine	61	none	88	rare
Leon	74	swimming	34	average

(c) **Conceptual** Name 3 hints that a decision tree for regression has overfit

4. **Understanding gradient boosting and boosting for classification** Write down the steps for boosting to solve

(a) Adaboost for finding the animal label

(b) Gradient boosting for finding the height of people

In both cases, use the tables given in the previous problems, and assume the weak learners are decision stumps.

After constructing the boosted trees, infer on the test set and report the classification error or mean squared error of regression.

5. **Another gradient boosting example** Suppose that we wish to fit $F(x_i) \approx y_i \in \mathbb{R}$ for $i = 1, ..., m$, using

$$F = \sum_{t=1}^{T} f_t, \qquad f_t \in \mathcal{H} \text{ some function class.}$$

I have a magic black box that can solve

$$f = \operatorname*{argmin}_{f \in \mathcal{H}} \sum_{i=1}^{m} (f(x_i) - u_i)^2$$

for any choice of $u_1, ..., u_m$.

Suppose that the function class is 1-sparse linear functions; that is,

$$\mathcal{H} = \{f_{k,\alpha} : f_{k,\alpha}(x) = ax[k],\ a \in \mathbb{R}\}.$$

Describe the first 3 steps of the gradient boosting algorithm, using this function class. (This method is a variant of something called *matching pursuit*.)

6. **Understanding Kmeans and GMM**

(a) I have a bunch sweaters: red, blue, and green. The amount of mothballs on each sweater forms a Gaussian distribution.

- Red sweaters have a distribution of mothballs with mean 10 and variance 1
- Blue sweaters have a distribution of mothballs with mean 1 and variance 100. (Assume "negative mothballs" are possible. Maybe they're butterflies.)
- Green sweaters have a distribution of mothballs with mean 100 and variance 14.

I have twice as many green sweaters as red sweaters, and I have an equal number of red and blue sweaters.

I pick up a sweater at random and it has 50 mothballs. What are the probabilities that the ball is red, blue, or green?

(b) 10 years later, I'm plagued with the same problem: 20 sweaters, all with mothballs, but they are different than the sweaters from the previous problem. I think the number of mothballs on the sweater has something to do with the sweater's material, but I do not know what the material is. The 20 sweaters have the following number of mothballs:

3,44,4,5,1,5,43,2,4,67,45,2,63,13,33,2,14,6,25,12.

Starting with a seed of 0, 10, 50, cluster these sweaters into 3 groups. Then run one episode of Kmeans and return the new cluster centers.

(c) Using the Gaussian mixture model from part (a), return the probability of each color of each of the 20 sweaters; that is, return $\pi_{i,k} = \mathbb{E}[\text{sweater i has color k}]$.

(d) Using the $\pi_{i,k}$ values from the previous section, return a new μ_1, μ_2, μ_3 and $\sigma_1, \sigma_2, \sigma_3$ and $\alpha_1, \alpha_2, \alpha_3$ to fit a slightly better Gaussian mixture model to this data. [1]

7. **Understanding PCA and matrix factorization**

(a) Consider the generalized matrix factorization problem

$$\underset{U,V}{\text{minimize}} \qquad f(UV^T).$$

Show that an infinite number of global minimizers U and V exist. (Hint: suppose that there exists at least one X where $f(X) = \min_{X'} f(X')$.) Note that this property does not require f to be convex.

(b) The Eckhart-Young-Mirsky theorem states that the solution to the problem

$$\begin{array}{ll} \underset{X}{\text{minimize}} & \|X - R\|_F \\ \text{subject to} & X \text{ has rank } r \end{array}$$

has a global minimizer

$$X = U\Sigma_r V^T$$

where $R = U\Sigma V^T$ is the SVD of F and Σ_r is Σ but with singular values that are not of the r largest values set to 0.

- Use this property to construct a solution to the symmetric matrix factorization problem

$$\underset{Z \in \mathbb{R}^{m \times r}}{\text{minimize}} \quad \|ZZ^T - R\|_F.$$

- Show that there are an infinite number of potential solutions Z and write down what they are.

(c) **Binary matrix factorization** After solving the matrix factorization problem, I receive a user and movie matrix of

$$U = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \qquad V = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

How many different "archetypes" of people and movies are there? (Correct answer is 2, how to argue this?)

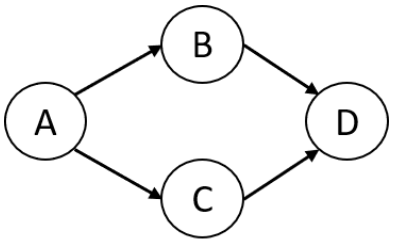
[1] This question involves a lot of numbers and probably requires a calculator or python or matlab. In the real test, I will simplify things so that you may still need a calculator, but not python or matlab. So, while you are free to use all these tools, avoid using any built-in fancy functions to answer this question.

(d) **Imputation** Given the following rating matrix

$$R = \begin{bmatrix} 1 & -1 & 1 & -1 \\ -1 & ? & -1 & 1 \\ 1 & 1 & ? & 1 \\ -1 & -1 & -1 & -1 \end{bmatrix}$$

and knowing that $R = UV^T$ where U and V both have 2 columns, fill in the values at the question marks with values 1 or -1.

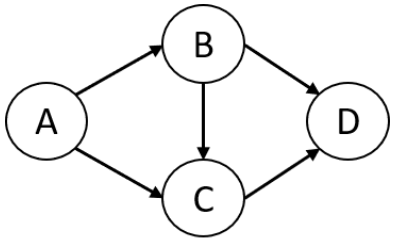
8. **Understanding directed graphical models** Consider the graphical model below, where each random variable A, B, C, D can take value either 0 or 1. Your goal is to infer $\mathbf{Pr}(D)$.



(a) What information do you need, based on this model, to get $\mathbf{Pr}(D)$?

(b) What can you tell me about $B|A$ (read: the event B given event A) and $C|A$?

(c) Now suppose we also add another directed arrow, as shown below. Is this still a DAG? What problems might arise when doing inference over this graphical model?



9. **Understanding multiclass logistic regression** Suppose that I solve the multiclass logistic regression problem, and reach the following conclusion, based on my 5 data points $X = [x_1, x_2, ..., x_5]^T$ and 3 classifier vectors $\Theta = [\theta_1, \theta_2, \theta_3]^T$:

$$X\Theta^T = \begin{bmatrix} 1 & -2 & 3 \\ -1 & 2 & -3 \\ 3 & -2 & 1 \\ 3 & 2 & -1 \\ -3 & -2 & 1 \end{bmatrix}$$

The following questions assume a multinomial logistic model.

(a) What is the learned probability that x_3 belongs in class $k = 2$?

(b) What is the log likelihood of the current setup?

(c) What is the gradient of $\log(\mathbf{Pr}(\hat{y}_i|y))$ given these current values?

10. **Teamwork.** Alice, Bob, and Carlos are all working on their problem set together. Each question is a True/False question. The group works by each guessing the answer, and writing down the majority vote answer.

(a) Given that each friend is independently 75% sure that the answer is True, what is the probability that they will write down the answer True?

(b) Given that each friend is independently $p\%$ sure that the answer is True, modeling true as +1 and false as -1, what is the mean and variance of their guess?

(c) The professor walks in. The three ask her for the answer. She mumbles "I think it's $\bar{s}$", where $\bar{s}$ is either True or False. But she hasn't been sleeping for the past 3 days so it's not certain that she even heard the question. The students decide to take their majority vote guess, s_{MLE}, and weight it with the professor's answer, guessing

$$\hat{s}_{\text{MAP}} = \textbf{sign}(\hat{s}_{\text{MLE}} + \rho \cdot \bar{s})$$

where a guess +1 means True and a guess -1 means False. What is the mean and variance of this new guess, as a function of ρ?

(d) While the three are good friends, the score overall is curved, so there is incentive for friends to sabotage each other, in order to get better scores. After doing some problems, Bob and Carlos start to suspect Alice of sabatoging. The sabateur strategy is a bit hard to detect: Alice will only sabatoge under 2 conditions

- she knows the right answer,
- Carlos and Bob disagree on the answer.

Then, with some nonzero probability, Alice will sabatoge by intentionally giving the wrong answer, making the group answer wrong, and then herself writing down the right answer.

Suppose you are given a table, such as

Correct answer	Alice	Bob	Carlos
True	True	False	False
False	True	True	False
False	False	True	True
True	False	True	True
True	False	True	True
⋮			

i. Propose a graphical model from which you can do some inference and return the probability that Alice is being malicious.

ii. Suppose you have access to the prior probabilities that Alice, Bob, or Carlos would know the right answer. Show how you would infer the probability that Alice is being malicious, given these readings.

11. **Bayes vs Asymptotic Nearest Neighbor** Harry Potter's twin sons Barry and Larry Potter have entered another year at Hogwarts, and there's a very nonzero probability that one or both of them will end up dead before graduation, as that school is very dangerous. In fact, the lifespan of a Hogwarts student can be modeled as an exponential distribution with mean λ_{House} years, where $\lambda_{\text{Gryffindor}} = 1$, $\lambda_{\text{Hufflepuff}} = 10$, $\lambda_{\text{Slytherin}} = 3$, $\lambda_{\text{Ravenclaw}} = 7$.

- What's the probability that Barry is still alive 5 years later, given that his chances of being sorted to any of the 4 Houses is equal?
- Suppose you know that Barry was sorted to either Gryffindor or Hufflepuff. You know that Barry died in year x. As a function of x, explain whether there is a higher probability that Barry was in Gryffindor or Hufflepuff.
- Sadly, both twins died at the exact same time. I know that Larry was sorted to Ravenclaw, so I just guess that Barry was in Ravenclaw as well. Describe a lower and upper bound on the risk that I was wrong.

12. **Dual and kernel SVMs**

- Derive the dual formulation of the soft and hard margin linear SVM problems, and show how to recover the primal solution from the dual solution.
- Using the dual kernel SVM formulation, show how the solution can be used to classify new points.
- Show that if a kernel matrix is not positive semidefinite, then the dual problem cannot be solved by minimizing a convex objective.

13. **Election season.** I am a professional polling analyst, and my job is to predict who the people of Suffolk county will vote for, in the year 3035. The options are

- vote for Candidate Martian,
- vote for Candidate Astroid Belt.

Suppose that there is no self-reporting bias; everyone I call answers my questions honestly. And, everyone is going to vote; no abstaining. And, assume that no one is conspiring with each other; each vote is given independently. There are 1.5 million people in Suffolk County.

I have now randomly polled 100 people, who have given me their voting reports, and 55 of the people I polled claimed they will vote for Candidate Astroid Belt, with 45 of people claiming favor for Candidate Martian.

(a) Give the maximum likelihood estimate of the true percent of people who will vote for Candidate Martian.

(b) How certain am I that this estimator is correct, within 5% error margin? (report in terms of m)

(c) **Dirty politics** It turns out that Candidate Astroid Belt is my favorite candidate, because she promised me $1,000,000 of bribes if she wins. On the other hand, Candidate Martian is going to raise my taxes, and I will lose $1,000.

i. What is my Bayes reward if I don't do anything and let the election play out as shown? What is my minimax reward?

ii. What is my certainty that Candidate Astroid Belt will win by a 1% margin?

(d) There's no more time left to do polling, but I have one more trick up my sleeve: I can stuff ballot boxes with fake tickets for candidate Astroid Belt, to help her win!

i. How many ballots do I need so that my Bayes Reward is more than $0?

ii. How many ballots do I need so that my minimax reward is more than $0?

iii. How many ballots do I need so that I can be 90% certain that my candidate will win by more than a 5% margin?

14. **Dual SVMs** Consider the following set of training data

i	x_i	y_i	$\phi(x_i)$
1	$(0, 2)$	-1	
2	$(-\sqrt{2}, \sqrt{2})$	-1	
3	$(0, 1)$	1	
4	$(\sqrt{2}, -\sqrt{2})$	-1	
5	$(\frac{-1}{\sqrt{2}}, (\frac{-1}{\sqrt{2}})$	1	
6	$(\frac{1}{\sqrt{2}}, (\frac{1}{\sqrt{2}})$	1	
7	$(1, 0)$	1	
8	$(\sqrt{2}, \sqrt{2})$	-1	
9	$(0, -2)$	-1	
10	$(2, 0)$	1	

(a) Plot the points

(b) For the following kernel functions, construct a kernel function $\phi : \mathbb{R}^2 \to \mathbb{R}$ that best separates the points. Hint: there exists one that can separate all but 1 point.

(c) Show that for this particular training set, using the kernel function you picked, compute analytically the solution of the soft-margin SVM problem.

Hint: try first simplifying (**??**) as much as possible, using what you computed for $\phi(x_i)$.

(d) Assuming that $\rho > 0$, does the value of ρ affect the classification accuracy? Why or why not?

(e) Assuming that $\rho > 0$, does the value of ρ affect the minimum margin distance of the resulting classifier? Why or why not?

(f) Do your answers change for different values of c?

(g) Solve the dual of the soft-margin SVM problem. A unique solution may not exist, but there should be enough information that the primal variable can be recovered.

(h) Show that if ρ was not specified, then the dual problem would be unbounded above for this particular data set, and argue that it is because the dataset is unseparable. (Hint: the dual is unbounded when the primal is infeasible.)

(i) Do some sensitivity analysis. When $c = 1$, it should be that for any positive value of ρ, the classifier will correctly classify all but one point. At what value of c is this no longer true? [2]

15. **Dimensionality reduction** (This exact kind of problem won't feature on the final, where there won't be any coding. But I do recommend trying some simple version of this, as it will help solidify the concepts that *will* be on the final!.) Pick a piece of data from the UCI machine learning repository, which is under some classification task.

    - Code up a linear classification tool, using sklearn or your own code.
    - Using the following dimensionality techniqes, experiment with how compressible the dataset is by seeing how small the data feature dimension can be while maintaining 90% performance. Do again for 99% performance.
        - PCA, MDS, JL-featured random projections, isomap, tsne
    - Code up a solver for a dual SVM. Using this construct, solve this problem using an RBF kernel. How does the computational complexity trade off with the performance gain?
    - Implement Nystrom sampling of the RBF kernel. Can the kernel be compressed enough so that we can regain low computational complexity but keep the performance gain?

16. Anything from the midterm practice exam is also fair game, though will not be the focus of the final.

17. Anything that was on a homework assignment (non-coding problem) is also fair game.

[2] The robustness against outliers is a feature of the hinge loss, which is a cornerstone of the soft-margin SVM problem.