CSE 512: Midterm review

---

1. **Nomenclature** Define the following terms
   (a) model vs loss function vs data
   (b) prior vs posterior vs likelihood
   (c) MLE vs MAP
   (d) Expected risk vs empirical risk
   (e) prediction, forecasting
   (f) PAC learning
   (g) biased vs unbiased estimator
   (h) Bayes risk, minimax risk

2. **Linear models for classification**
   (a) Given the following datasets, would you propose a linear, generalized linear, or neither model to predict y given x? Jusify your answer.

i.

Temperature (x)	10°F	25°F	50°F	70°F	100°F
Wear a sweater? (y)	yes (+1)	yes (+1)	no (-1)	no (-1)	no (-1)

ii.

Time of day (x)	7h00	10h00	12h00	15h00	20h00
Wear a sweater? (y)	yes (+1)	no (-1)	no (-1)	no (-1)	yes (+1)

iii.

Chance of earthquake (x)	0.01%	0.1%	1%	10%	100%
Wear a sweater? (y)	yes (+1)	no (-1)	yes (+1)	no (-1)	no (-1)

iv.

Commute speed (km/hr) (x)	walking (1)	jogging (5)	biking (20)	horse and buggy (30)	driving (60)
Wear a sweater? (y)	yes (+1)	yes (-1)	yes (+1)	no (-1)	no (-1)

   (b) For the task of predicting whether you should wear a sweater, using the features from the previous problem, propose a way of using logistic regression to construct a model. Construct some fake data for yourself, train the model, and use it to predict whether you should wear a sweater in the following scenarios
      i. Warm day 70°F, early morning 8h00, no chance of earthquake (0%), bringing the horse and buggy
      ii. Cold day 20°F, late afternoon 16h00, medium chance of earthquake (25%), jogging

3. **Linear regression**
   (a) Describe a convex loss function where, when minimized, returns θ that maximizes the likelihood of the following model:

$$y_i - x_i^T\theta \sim \mathcal{N}(\mu_1, \sigma_1), \qquad \theta \sim \mathcal{N}(\mu_2, \Sigma_2)$$

   where μ_1, σ_1 are scalars, μ_2 is a vector, and Σ_2 is a PSD matrix; all 4 of these constants are known.
   (b) A is a symmetric positive semidefinite matrix, and the condition number of A is $\bar{\kappa}$. The maximum eigenvalue of A is $\lambda_{\max}$. Write an expression $\kappa(\rho)$ that returns the condition number of $A + \rho I$.
   (c) $B = XX^T$ and the singular value decomposition of X is

$$X = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{2}} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

   Write an expression $\kappa(\rho)$ that returns the condition number of $B + \rho I$. What happens when $\rho = 0$?

(d) $C = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. Write an expression $\kappa(\rho)$ that returns the condition number of $C + \rho I$

(e) Consider the linear regression problem

$$\underset{x}{\text{minimize}} \quad f(x) = \|Ax - b\|_2^2 + \rho\|x\|_2^2$$

i. First consider $\rho = 0$. What are the normal equations? Write them down.

ii. Now consider general ρ. Write down the normal equations again, and describe how they relate to the gradient of the objective.

iii. Describe how you would solve for the solution x when

A. A is a wide matrix (more columns than rows) and $\rho = 0$

B. A is a tall matrix (more rows than columns) and has full column rank, and $\rho = 0$

iv. Are either case made easier of $\rho > 0$?

v. Write down the condition number for $f(x)$.

(f) Suppose that I was given a set of feature/label pairs x_i, y_i for $i = 1, ..., m$, and I wish to do linear regression to find a model that, for a new vector x, well-approximates its corresponding value y.

But, for hardware reasons, I cannot hold onto the values x_i; instead, I hold onto the Fourier transform of x_i; that is, II have access to $u_i = Fx_i$ where F is the DFT matrix.

The DFT matrix is in general super efficient to apply (it doesn't really require a full matrix-vector multiplication). Additionally, it is a unitary matrix, so that $FF^T = F^TF = I$.

i. Describe the inverse DFT matrix, e.g. what does F^{-1} look like?

ii. Write down the least squares system we need to solve such that we retrieve $u = Fx$, where $Ax \approx b$. This system should look like

$$\underset{u}{\text{minimize}} \quad \|\hat{A}u - \hat{b}\|_2^2.$$

What are $\hat{A}$ and $\hat{b}$?

iii. Given κ the condition number of A^TA, what is the condition number of $\hat{A}^T\hat{A}$?

4. **Binary classification**

(a) Consider the following dataset

$x[1]$	$x[2]$	y
-1.0	-1.0	+1
-0.25	2.0	-1
-2.0	-0.25	+1
-0.5	0.5	-1
0.5	-1.25	-1
0.25	2.0	+1
3.0	-0.25	-1
2.5	1.0	+1

i. In words, can you describe the rule being used to generate y from $x \in \mathbb{R}^2$? (Hint: start by plotting the points)

ii. Is this dataset linearly separable? Why or why not?

iii. Propose a generalized linear model using 2-order polynomials (where the highest degree is 2) that can separate this data. For this model, compute the margin for each datapoint and report the minimum margin.

5. **Gradient descent** Consider the following generalized loss function

$$f(\theta) = \frac{1}{m}\sum_{i=1}^{m} g(y_i x_i^T \theta)$$

where $g : \mathbb{R} \to \mathbb{R}$ is convex and differentiable everywhere, and $\theta \in \mathbb{R}^n$.

(a) What is the gradient and Hessian of f? What are their dimensions?

(b) Would f be convex if g were not convex? If g were concave? Why or why not?

(c) What would be good qualities to impose on g such that minimizing f produces a margin maximizing classifier?

(d) Now consider $F(\theta) = f(\theta) + \rho\|\theta\|_2^2$ for some $\rho > 0$. What is the condition number of F (in terms of properties of x_i, y_i, and g?

(e) Write out pseudocode implementing gradient descent for both the regularized and unregularized form. Specifically, fill in the gaps, and include how I would go about computing a step size given y_i, x_i, and g.

```
def grad_desc_unreg(X,y):


    <fill me in >


    return theta


def grad_desc_reg(X,y,rho):


    <fill me in >


    return theta
```

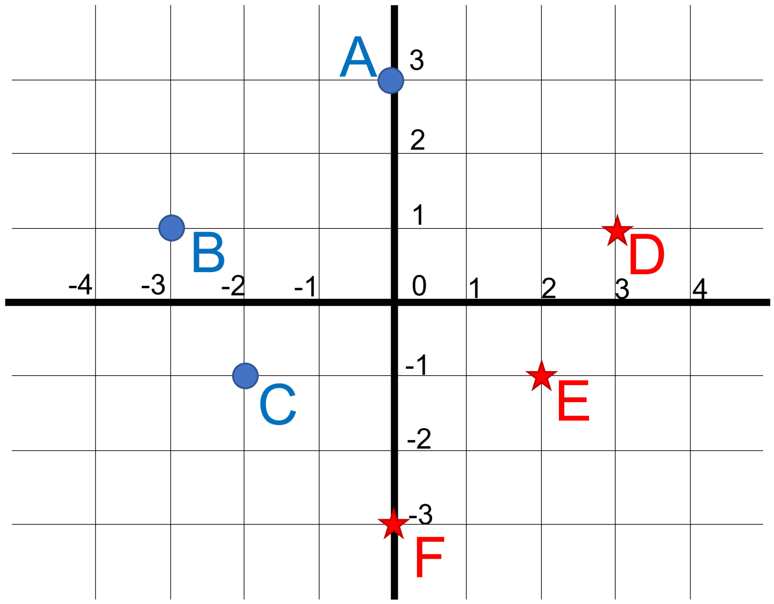
Assume that you have access to functions on g, namely

```
def g(theta):
    <computes z = g(theta)>
    return z

def g_grad(theta):
    <computes zp = g'(theta)>
    return zp

def g_hess(theta):
    <computes zpp = g''(theta)>
    return zpp
```

6. **Margins.** I have 6 datapoints, plotted below.



You should interpret each feature vector as

$$x_A = \begin{bmatrix} 0 \\ 3 \end{bmatrix}, \qquad x_B = \begin{bmatrix} -3 \\ 1 \end{bmatrix}, \qquad x_C = \begin{bmatrix} -2 \\ -1 \end{bmatrix}, \qquad x_D = \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \qquad x_E = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \qquad x_F = \begin{bmatrix} -3 \\ 0 \end{bmatrix}.$$

The labels correspond to blue points (-1) and red points (+1), e.g.

$$y_A = y_B = y_C = -1, \qquad y_D = y_E = y_F = 1.$$

(a) **Draw some decision boundaries.** On the plot above, draw a line (solid) corresponding to the set

$$\mathcal{S}_1 = \{x : x^T\theta_1 = 0\}$$

where $\theta_1 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$. Also, draw a line (dashed) corresponding to the set

$$\mathcal{S}_2 = \{x : x^T\theta_2 = 0\}$$

where $\theta_2 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$.

(b) Fill in the table below with each *datapoint's* margin, e.g. the distance from each feature vector to the margin:

	dist to $\mathcal{S}_1$	dist to $\mathcal{S}_2$
A		
B		
C		
D		
E		
F		

(c) I use the usual linear predictor to deal with new points:

$$y = \textbf{sign}(\theta^T x)$$

i. If I pick $\theta = \theta_1$, which points (A,B,C,D,E,F) are my *support vectors*? What is this *predictor's* minimum margin?

ii. If I pick $\theta = \theta_2$, which points (A,B,C,D,E,F) are my *support vectors*? What is this *predictor's* minimum margin?

iii. Which choice of θ maximizes the minimum margin?

(d) Argue that $\theta = \theta_1$ is in fact the optimal margin maximizing choice. Do this in two steps:

i. Argue that changing the norm $\|\theta\|_2$ does not affect the minimum margin.

ii. Argue that changing the rotation of θ, (e.g. $\theta_{[1]}/\theta_{[2]}$) will always reduce the margin to one or more of the support vectors of θ_1.

7. **Support vector machines** A common depiction of the SVM problem formulation is

$$\begin{array}{ll} \underset{\theta\in\mathbb{R}^n, s\in\mathbb{R}^m}{\text{minimize}} & \|\theta\|_2^2 + \lambda \sum_{i=1}^m \max\{0, s_i\} \\ \text{subject to} & y_i x_i^T \theta = 1 - s_i \end{array} \tag{1}$$

Suppose I solve (1) and I receive some optimal solutions for θ^*, s^*.

(a) What is the margin of the classifiers I received?

(b) Suppose $m = 100$ and $n = 25$. For my values of s, 34 of them are nonzero, and 66 of them are 0. What is an upper bound on how many misclassified training points there are?

(c) Would you expect my number of misclassified points to increase or decrease if I increase/decrease λ?

(d) A cat walks across my keyboard and accidentally deletes θ^*. Am I screwed? Can I recover θ^*?

8. **Convex sets** Decide if the following sets are convex. Either prove they are, or provide a counterexample to show they are not.

(a) $\mathbf{null}(A) = \{x : Ax = 0\}$ for some matrix A

(b) The set of intervals $[a, b] = \{x : a \leq x \leq b\}$

(c) The set of vectors $\{x \in \mathbb{R}^n : \prod_i x[i] = 0\}$

9. **Convex functions** Decide if the following functions are convex. Either prove they are, or provide a counterexample to show they are not.

(a) $f(x) = x^p$ for $p = 1, 2, 3, ...$

(b) $f(x) = x^p$ for $p = 1, 2, 3, ...$ over the domain $x \geq 0$

(c) $f(x) = \sqrt{|x|}$

(d) $f(x) = \sigma(x)$ where $\sigma(x) = 1/(1 + e^{-x})$

(e) $f(x) = \log(x)$

(f) $f(x) = \log(\sigma(x))$ (for same definition of σ)

(g) $f(x) = \mathbb{E}_y[f(x, y)]$ where f is convex over x, for fixed y

10. **Optimality.** Classify x^* as a local min, local max, global min, global max, saddle point, stationary point, or none. (More than one may apply.) Justify your answer

(a) $f(x) = x^2$ and $f'(x) = 0$

(b) $f(x)$ is convex and $f'(x) = 0$

(c) $-f(x)$ is convex and $f'(x) = 0$

11. **Point estimation** (Use of simple calculator permitted.) Recall Hoeffding's inequality

$$\mathbf{Pr}\left(\frac{1}{m}\sum_{i=1}^m x_i - \mathbb{E}[X] \geq \epsilon\right) \leq \exp(-2m\epsilon^2).$$

Suppose I have a fair coin (50% chance of getting heads or tails) and I flip the coin m times.

(a) How many flips until I am 90% certain that between 25%-75% flips are heads?

(b) If I flip the coin 100 times, how certain am I that between 45 to 55 flips are tails?

(c) My boss does not believe that the coin is fair, and tells me to keep flipping the coin until I am 1% certain of whatever I report. I flip until I get carpal tunnel syndrome, which is about 234 flips, of which 113 are heads. What range of values for heads/tails can I report and still guarantee 99% certainty the real weighting of the coin?

12. **Biased or unbiased?** In the previous homework, you worked with the exponential distribution, defined by

$$p_\lambda(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{else.} \end{cases}$$

In particular, you showed that, given samples $x_1, ..., x_m$ drawn i.i.d. from this distribution, that $\hat{\theta} = \frac{1}{m}\sum_{i=1}^m x_i$ served as both the maximum likelihood and unbiased estimator for the true mean, $\frac{1}{\lambda}$.

(a) Derive the MLE for λ.

(b) Show that this MLE is biased. Hint: $f(x) = 1/x$ is a strictly convex function whenever $x > 0$.

13. **Estimation and Risk analysis**

I am a stock broker, and on my computer screen, I see 100 stocks. At each given day, the stocks report a return rate (e.g. if I had invested \$x in stock i, my share of that stock at the end of that day would be worth $(1 + r_i)\$x$. I record these rates over the past 100 days, and notice that they can be modeled well by a Gaussian distribution, with mean μ_i and standard deviation σ_i for the ith stock.

(a) **Even investment** Let's assume that I decide to diversify my funds to the limit, and give an equal amount of money (say, \$10) to every stock.

i. What is the maximum likelihood estimate for my expected return that day?

ii. I decide to estimate the standard deviation of my return by using $\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{100} \sigma_i^2}{100}}$. Is this a good idea? bad idea? Is it a maximum likelihood estimator? of what? Is it biased/unbiased?

(b) **Consulting** I have two clients, Alice and Bob. They both want me to recommend one stock where they will put all their funds.

i. Using the rate of money earned each day as the reward function, what is the Bayes reward for picking stock i?

ii. Using the rate of money earned each day as the reward function, what is the Minimax reward for picking stock i?

iii. Alice is representing a business with tons of insurance protection. Her job is to earn as much money in the long run. Should I recommend to her the fund that maximizes the Bayes reward or the minimax reward?

iv. Bob is investing his life savings. He really can't afford to lose a single cent. Should I recommend to her the fund that maximizes the Bayes reward or the minimax reward?

(c) **Diversification** Suppose that on the next computer screen, there are 10 additional stocks. Each stock, on each day, has a 25% chance of doubling your investment, and a 75% chance of losing half your money. I have \$100 to invest.

i. What is your Bayes risk if you invest all your money in one stock?

ii. What is your Bayes risk if you invest your money evenly in every stock?

iii. Using Hoeffding's inequality, what are the chances that you will lose half or more of your money in one day if you invest in 1 stock? 10 stocks? What if there were 100 such stocks?
Hint: It's helpful to first ask yourself if there is an implicit bound on how much money you can earn in one day.

14. **Nearest Neighbors and Bayes for regression** Suppose I want to sell my house. It's a beautiful 2 story blue house made primarily of wood, built in 1980, and located in Martha's Vinyard. I want to price this house reasonably, but I'm not really sure what a good price will be. I look around and see the following other houses

	# stories	color	construction material	distance to my house	built year	sold at price
Alice's house	2	pink	wood	5 miles	2016	$ 1 million
Benjamin's house	4	green	straw	next door	1776	$100
Carly's house	1	brown	concrete	2.5 miles	1980	$ 300,000
Dasha's house	3	white	wood	100 miles	1999	$150,000
Elliot's house	1	purple	brick	1800 miles	2010	$500,000

(a) Discuss how you would use a KNN regression model to pick a good price for my house. That is, design a reasonable distance function, compute the "distances" to the features of each friend's house, decide on a reasonable value for K, and give a reasonable price.

(b) Now assume that I have $m \to +\infty$ friends. In this regime, I have at least one friend whose house is in the same location as mine and has basically the exact same features. That friend sold their house for $25 million. Does this mean that I am guaranteed to also sell it for this price? Why or why not?

(c) Suppose that house prices were dependent only on color and construction material. Discuss how you could form a Bayes and naive Bayes regression model to figure out the maximum likelihood price that I should use. How much training data would you need if there are 5 possible colors, 5 distinct types of construction material, the true price is between 0 and $ million, and I want to be accurate up to the nearest $100?