STAT 154/254 Final

Raymond Tsao

TOTAL POINTS

105 / 148

QUESTION 1

Multiple Choice Questions 80 pts

1.1 **4 / 4**

✓ **+ 4 pts** *Correct*

+ 0 pts Incorrect

1.2 **4 / 4**

✓ **+ 4 pts** *Correct*

+ 0 pts Incorrect

1.3 **4 / 4**

✓ **+ 4 pts** *Correct*

+ 0 pts Incorrect

1.4 **4 / 4**

✓ **+ 4 pts** *Correct*

+ 0 pts Incorrect

1.5 **4 / 4**

✓ **+ 4 pts** *Correct*

+ 0 pts Incorrect

1.6 **0 / 4**

+ 4 pts Correct

✓ **+ 0 pts** *Incorrect*

1.7 **4 / 4**

✓ **+ 4 pts** *Correct*

+ 0 pts Incorrect

1.8 **0 / 4**

+ 4 pts Correct

✓ **+ 0 pts** *Incorrect*

1.9 **4 / 4**

✓ **+ 4 pts** *Correct*

+ 0 pts Incorrect

1.10 **4 / 4**

✓ **+ 4 pts** *Correct*

+ 0 pts Incorrect

1.11 **4 / 4**

✓ **+ 4 pts** *Correct*

+ 0 pts Incorrect

1.12 **0 / 4**

+ 4 pts Correct

✓ **+ 0 pts** *Incorrect*

1.13 **0 / 4**

+ 4 pts Correct

✓ **+ 0 pts** *Incorrect*

1.14 **4 / 4**

✓ **+ 4 pts** *Correct*

+ 0 pts Incorrect

1.15 **0 / 4**

+ 4 pts Correct

✓ **+ 0 pts** *Incorrect*

1.16 **4 / 4**

✓ **+ 4 pts** *Correct*

+ 0 pts Incorrect

1.17 **0 / 4**

+ 4 pts Correct

✓ **+ 0 pts** *Incorrect*

1.18 **4 / 4**

✓ **+ 4 pts** *Correct*

+ 0 pts Incorrect

1.19 **4 / 4**

✓ **+ 4 pts** *Correct*

+ 0 pts Incorrect

1.20 **4 / 4**

✓ **+ 4 pts** *Correct*

+ 0 pts Incorrect

QUESTION 2

2 OLS vs PCA **13 / 13**

✓ **+ 3 pts** *Part (a) correct*

✓ **+ 6 pts** *Part (b) correct*

✓ **+ 4 pts** *Correct illustration for part (c)*

QUESTION 3

3 Decision Tree Construction **5 / 7**

- 0 pts Correct

✓ **- 1 pts** *Split corresponding to t_4 is used but not needed.*

- 7 pts Incorrect / no attempt

✓ **- 2 pts** *Average not used*

+ 1 *Point adjustment*

QUESTION 4

4 Laplace Linear Regression **9 / 9**

✓ **- 0 pts** *Correct*

- 9 pts No progress / incorrect.

- 2 pts Insufficient comparison with OLS.

- 5 pts No answer for (b) or incorrect.

QUESTION 5

5 VC Dimension **2 / 8**

+ 8 pts Correct

+ 7 pts Correct, but equality case missed

+ 4 pts Wrong logic but right ideas.

+ 0 pts Blank / no progress / completely incorrect steps.

+ 2 *Point adjustment*

QUESTION 6

6 Margin Dynamics in Boosting **2 / 9**

+ 5 pts a) is fully correct

+ 4 pts a) has little problems

+ 2 pts a) is partially correct

✓ **+ 0 pts** *a) is incorrect*

+ 4 pts b) is correct

✓ **+ 2 pts** *b) is partially correct*

+ 0 pts b) is incorrect

QUESTION 7

7 Convolutional Neural Net **10 / 10**

✓ **+ 2 pts** *a) Correct*

✓ **+ 2 pts** *b) Correct*

✓ **+ 2 pts** *c) Correct*

✓ **+ 2 pts** *d) Correct*

✓ **+ 2 pts** *e) Correct*

+ 0 pts Incorrect/missing computations

QUESTION 8

8 Lloyd's algorithm **8 / 12**

✓ **+ 6 pts** *a) Correct*

+ 4 pts a) Correct with some problems

+ 2 pts a) Partially correct

+ 6 pts b) Correct

+ 4 pts b) Correct with some problems

✓ **+ 2 pts** *b) Partially correct*

+ 0 pts Incorrect/ No answer

In b) the convergence of the algorithm is influenced by the criteria you use for stopping it (otherwise we always converge to a loc minima). For example, if you halt the algorithm when there are no further changes in cluster assignments, it is possible to demonstrate that convergence is not guaranteed. This especially holds true if the method for assigning clusters does not include a definitive way to resolve ties when multiple centroids are equidistant.

Stat154/254 Modern Statistical Prediction and Machine Learning, Fall 2023
Final Exam

Instructor: Nikita Zhivotovskiy

GSI: Mriganka Basu Roy Chowdhury

Maximum Points: 148 points

Duration: 2 hours 50 minutes.

Write your name: Raymond Tsao

SID Number: 3037860126

Please select the course you are enrolled in

☑ STAT 154. ☐ STAT 254.

Exam Information and Instructions:

- This is a closed book, closed notes exam. The use of electronic devices is not permitted.
- The exam has a "list of useful" facts that can help you with some problems.
- Write any work you want graded on the front of each page. The reverse side of the page can be used as scratch paper. Additionally, write your SID number in the top right corner on every **extra** page you might need.
- Provide reasoning for all **Written Questions**. Please, write as clearly as possible.
- For multiple answer questions, fill in the answer for **ALL** correct choices. At least one of the answers is always correct. **NO** partial credit for the multiple answer questions.
- Please write your answers as clearly and legibly as possible.

I certify that all materials in the enclosed exam are my own original work and I followed Berkeley Honor Code.

Sign your name: Raymond Tsao .

Good Luck!

List of useful facts

SVD. Any rectangular m by n matrix A can be written as $A = U\Sigma V^T$, where U is a m by m orthogonal matrix; V is a n by n orthogonal matrix, Σ is a m by n rectangular diagonal matrix with nonnegative entries. The left-singular vectors of A are a set of orthonormal eigenvectors of AA^T. The right-singular vectors of A are a set of orthonormal eigenvectors of A^TA. The non-zero singular values of A (found on the diagonal entries of Σ) are the square roots of the non-zero eigenvalues of both A^TA and AA^T.

Eckart–Young–Mirsky theorem. Given an m by n matrix A, the truncated SVD A_p, for $p < \min\{m, n\}$, satisfies that for any matrix B of rank at most p,

$$\|A - A_p\|_F \leq \|A - B\|_F.$$

Here, $\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2}$ is the Frobenius norm of A.

VC dimension. Given a set of classifiers $\mathcal{F}$, defined on a domain $\mathcal{X}$, the VC dimension of $\mathcal{F}$ is the largest integer d for which there exists a subset $x_1, \ldots, x_d$ of $\mathcal{X}$ shattered by $\mathcal{F}$. To say that $\mathcal{F}$ shatters a subset $x_1, \ldots, x_d$ means that for every possible combination of binary labels for the d points, there exists at least one classifier in $\mathcal{F}$ that can realize this combination.

Convex Functions. A function $\phi : \mathbb{R}^p \to \mathbb{R}$ is convex if $\phi(\alpha x + (1-\alpha)y) \leq \alpha\phi(x) + (1-\alpha)\phi(y)$, for all x, $y \in \mathbb{R}^p$ and $\alpha \in [0, 1]$.

ROC curve. The ROC curve plots TPR against FPR. The Area Under the ROC Curve (AUC) is an area under the ROC curve and takes values in $[0, 1]$.

Vapnik-Chervonenkis Theory. Given a set of classifiers $\mathcal{F}$ with VC dimension d, the theory shows that for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, the inequality

$$|R(f) - R_N(f)| \leq 4\sqrt{\frac{2d\log(16N) + 2\log(1/\delta)}{N}},$$

holds simultaneously for all $f \in \mathcal{F}$. Here, $R(f)$ is the true risk (population/or expected test sample error) and $R_N(f)$ is the empirical risk based on a sample of size N. This result is significant as it bounds the difference between the true risk and the empirical risk, a measure of the generalization error, irrespective of the distribution of the data, based solely on the VC dimension and the size of the training set.

Entropy. The entropy of a dataset S is defined as:

$$H(S) = -\sum_{i=1}^{k} p_i \log_2(p_i),$$

where k is the number of classes, and p_i is the proportion of the instances that belong to class i in S. The information gain IG for a dataset S split on an attribute A with values $\{a_1, a_2, \ldots, a_m\}$ is given by:

$$IG(S, A) = H(S) - \sum_{j=1}^{m} \frac{|S_{a_j}|}{|S|} H(S_{a_j}),$$

where $|S|$ is the total number of instances in S, S_{a_j} is the subset of S where attribute A has the value a_j, $|S_{a_j}|$ is the number of instances in S_{a_j}, and $H(S_{a_j})$ is the entropy of S_{a_j}.

Sigmoid Function and Universal Approximation Theorem. A function $\sigma(z)$ is defined as sigmoidal if it satisfies the following limits: $\lim_{z\to-\infty} \sigma(z) = 0$ and $\lim_{z\to+\infty} \sigma(z) = 1$.

Theorem: Cybenko (1989). Given $\sigma(z)$ as a non-constant sigmoid function, for any continuous function $f(x)$ on the hypercube $[0, 1]^p$, there exist parameters H, $\alpha_h \in \mathbb{R}$, $w_h \in \mathbb{R}^p$, $w_0 \in \mathbb{R}$, such that a two-layer neural network can be constructed as follows:

$$\widehat{f}(x) = \sum_{h=1}^{H} \alpha_h \sigma(\langle x, w_h \rangle - w_0).$$

This network approximates $f(x)$ with any pre-defined accuracy ε, i.e.,

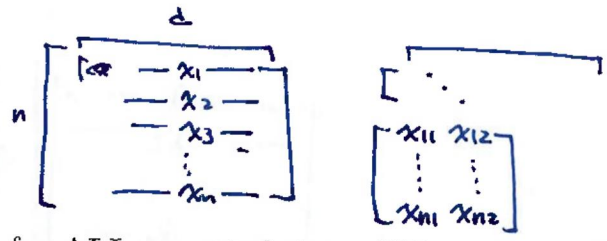
$$|\widehat{f}(x) - f(x)| < \varepsilon, \quad \forall x \in [0, 1]^p.$$

Multi-armed Bandits. In the context of multi-armed bandits, a gambler faces a slot machine with N arms, each with its own probability distribution D_i over the interval $[0, 1]$. The gambler's objective is to minimize the expected regret by sequentially choosing which arm to pull. The expected loss for each arm i is denoted by μ_i, and the arm with the optimal expected value is identified by $\tilde{I} = \arg\min_{1\le j\le N} \mu_j$. In each round t, the gambler selects an arm I_t, leading to a loss $X_{i,t}$ which is a random variable from D_i. The cumulative expected regret up to round T is defined as the sum of the differences between the loss incurred by the arm selected in each round and the loss that would have been incurred by the optimal arm, mathematically expressed as:

$$\text{E-regret} = \mathbb{E}\left(\sum_{t=1}^{T} \left(X_{I_t,t} - X_{\tilde{I},t}\right)\right),$$

where the expectation is taken over the distributions $\{D_i\}_{i=1}^{N}$ and the randomness in the arm selection process.

1. Multiple Answer Questions



For multiple answer questions, fill in the answer for **ALL** correct choices. **NO** partial credit on multiple answer questions. Multiple choice questions are all worth 4 **points.**

$\Sigma = X^T X$

2. Given a design matrix $X \in \mathbb{R}^{n \times d}$ and the corresponding vector of responses $Y \in \mathbb{R}^n$ representing n sample points with d features, you compute the sample covariance matrix Σ of your dataset and find that its rank is less than d. Which of the following statements is true?

- ■ At least one feature is a linear combination of others.
- ☐ All sample points in the dataset lie perfectly on a straight line in the d-dimensional space. (I didn't choose this, unselect.)
- ☐ The matrix Σ is invertible.
- ■ The least squares has infinitely many solutions.

3. Which of the following is a characteristic of ensemble learning methods?

- ☐ Ensemble methods require the use of base predictors that are exactly the same.
- ■ Ensemble methods can help in reducing both bias and variance, compared to individual models.
- ■ Ensemble methods combine multiple machine learning models to obtain better predictive performance.
- ☐ Ensemble methods cannot be used for classification tasks.

TP

TN

Type I:

False → True.

4. In the context of a binary classification task in machine learning, which of the following statements about True Positive (TP), False Positive (FP), False Negative (FN), True Negative (TN), True Positive Rate (TPR), and False Positive Rate (FPR) are correct?

negwere classified as positive.

- ■ (TP) refers to instances where the model correctly predicts the positive class.
- ■ (FN) occurs when the model incorrectly identifies a positive class as negative.
- ■ (FP) occurs when the model incorrectly predicts the negative class as positive.
- ☐ (TNR) is the proportion of actual negatives that are correctly identified, which is the same as (FPR).

5. Evaluate the following statements regarding whether certain machine learning models are examples of Generalized Linear Models (GLM) or not. Which of these models align with the principles of GLM?

- ☐ Support Vector Machines (SVM) are a type of GLM, as they use linear decision boundaries in their basic form.
- ■ Logistic Regression is a GLM, as it uses a logistic function to link the linear predictors to a binary outcome.
- ☐ The Perceptron algorithm is a GLM, since it involves linear combinations of input features to make predictions.
- ☐ Decision Trees can be considered a type of GLM because they implement linear splits of features at each decision node.

6. What is true for PCA?

- [x] The sum of the eigenvalues of the covariance matrix equals the total variance in the original data set.
- [x] The eigenvectors of the sample covariance matrix define the directions of maximum variance, known as principal components.
- [x] In PCA, the first principal component is the direction that maximizes the variance of the data.
- [x] PCA transforms the original variables into a new set of variables, which are the principal components, ranked in order of the amount of original variance they explain.

7. Evaluate the total regret in three different multi-armed bandit setups with the same average loss relative to the best arm. In Case (1), there is one arm ($K = 1$), in Case (2), there are ten arms ($K = 10$), and in Case (3), there are also ten arms ($K = 10$) but pulling any arm j also reveals the rewards of two neighboring arms ($j - 1$ and $j + 1$, with arms 1 and 10 as neighbors).

- [ ] The total regret in Case (1) might be non-zero, despite having only a single arm.
- [x] The standard Upper Confidence Bound (UCB) algorithm may not yield the same regret for Case (2) and Case (3).
- [x] Case (2) might incur a higher total regret compared to Case (3) if the extra information from neighboring arms in Case (3) is used effectively.
- [ ] Case (3) could have lower total regret than Case (1).

8. What advantage does using a random forest offer over a single decision tree?

- [x] Reduces overfitting by averaging multiple tree outputs.
- [x] Increases model accuracy through ensemble diversity.
- [ ] Simplifies the model decision rules for easier interpretation.
- [x] Improves predictive performance on unseen data.

8. Given a sequence of observations x_i with associated targets y_i, and the loss function $\ell(\theta, x_i, y_i)$ for $i = 1, \ldots, t$, identify the update rule for stochastic gradient descent using the learning rate η:

- [ ] $\theta_{t+1} \leftarrow \theta_t - \eta \nabla_\theta \ell(\theta_t, x_t, y_t)$,
- [x] $\theta_{t+1} \leftarrow \theta_t - \eta \sum_{i=1}^{t} \nabla_\theta \ell(\theta_t, x_i, y_i)$,
- [ ] $\theta_{t+1} \leftarrow \theta_t - \eta \nabla_{x_i} \ell(\theta_t, x_t, y_t)$,
- [ ] $\theta_{t+1} \leftarrow \theta_t - \eta \sum_{i=1}^{t} \nabla_{x_i} \ell(\theta_t, x_i, y_i)$.

9. Regarding Cybenko's theorem and neural networks, select the correct statements

- [x] A network with single hidden layer can approximate (almost) any function, minimizing the need for depth.
- [x] The theorem focuses on single-layer, sigmoidal networks.
- [x] Cybenko's theorem remains silent on the applicability to activations like ReLU.
- [x] It suggests width can substitute for depth, but doesn't rule out benefits of deeper networks.

10. In training a neural network with backpropagation and a single hidden layer to minimize a loss function $\ell(\theta)$, what does backpropagation primarily compute?

☐ Starts with the gradient of ℓ at the input layer, propagating through the network.

☐ Begins with gradients of activation functions in the hidden layer, affecting the entire network.

■ Computes the gradient of ℓ across all weights, starting from the output layer.

☐ Focuses only on adjustments to output layer weights.

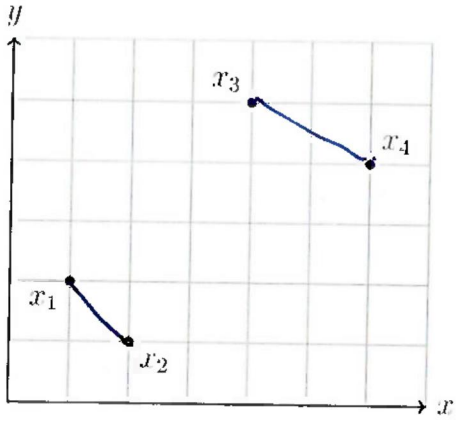
11. Consider a dataset with four points in a 2-dimensional space: $x_1 = (1, 2)$, $x_2 = (2, 1)$, $x_3 = (4, 5)$, $x_4 = (6, 4)$. After performing the first **two** steps of hierarchical clustering using the single-linkage method (distance between clusters is the distance between their closest members) with Euclidean distance as the metric, what is the cluster structure?

☐ $\{x_1\}$, $\{x_2\}$, $\{x_3, x_4\}$,

☐ $\{x_1, x_2, x_3\}$, $\{x_4\}$,

■ $\{x_1, x_2\}$, $\{x_3, x_4\}$,

☐ $\{x_1, x_2\}$, $\{x_3\}$, $\{x_4\}$.



12. The Bayes risk for a decision problem is zero under which conditions?

☐ Every feature in the dataset is independent of the class label.

☐ The class conditional probabilities, $P(X|Y)$, are identical for all classes.

■ The data points from different classes are perfectly distinguishable with no overlap in their distributions.

■ The prior probability of one class is overwhelmingly dominant, approaching 1, making other classes negligible.

13. Assess the correctness of the following statements about AdaBoost:

☐ When used with decision tree classifiers, tends to overfit in low-noise data environments.

☐ After each iteration, reweights each data point, decreasing the weights of misclassified instances.

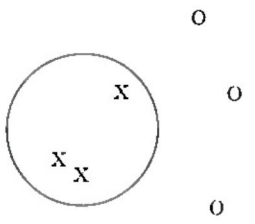
■ The algorithm shows decreased effectiveness in highly imbalanced classification scenarios.

☐ Its effectiveness in variance reduction is compromised when using strong learners as the base classifiers.

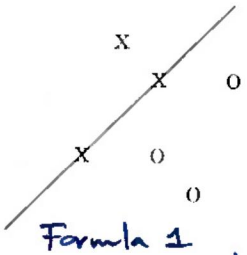
14. Match each of the following SVM kernel formulas with the correct graphical data separation representation: Formula 1 ($K(x, x') = x^T x'$). Formula 2 ($K(x, x') = (\gamma x^T x' + r)^d$). Formula 3 ($K(x, x') = \exp(-\gamma \|x - x'\|^2)$). Select the correct option:

- [x] Formula 1 - Picture 2, Formula 2 - Picture 3, Formula 3 - Picture 1.
- [ ] Formula 1 - Picture 1, Formula 2 - Picture 2, Formula 3 - Picture 3.
- [ ] Formula 1 - Picture 3, Formula 2 - Picture 1, Formula 3 - Picture 2.
- [ ] Formula 1 - Picture 3, Formula 2 - Picture 2, Formula 3 - Picture 1.

Picture 1:

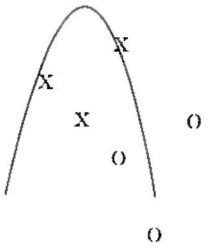


Picture 2:



Formula 1 linear kernel.

Picture 3:



15. Evaluate the accuracy of the following statements about AlexNet:

- [x] AlexNet was the first architecture to introduce Convolutional Neural Networks (CNNs) to the field of deep learning.
- [x] AlexNet was instrumental in reducing the error rate significantly in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012.
- [ ] It utilized the linear function as the activation function.
- [ ] The architecture of AlexNet is notable for its early use of the attention mechanism to enhance feature extraction.

16. In applying the spectral clustering algorithm to a dataset, the Gaussian (Radial Basis Function, RBF) kernel is used as the similarity measure to construct the similarity matrix. After computing the graph Laplacian, the following eigenvalues (in ascending order) are obtained:

$$0, 0.05, 0.07, 0.1, 1.5, 1.6, 1.8, 2.0, 2.1, 2.2, 2.3, \ldots, 100.0.$$

Based on these eigenvalues, what is the most natural number of connected components (clusters) for this dataset?

- [ ] 2
- [x] 3

assuming that 0.1 is too large from zero so that it cannot be treated as a clus

- [x] 4
- [ ] 8

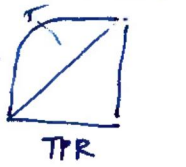
I assumed that 0.1 is closed enough to 0 so that it can be treated as a cluster (especially there's a huge jump from 0.1 to 1.5)

17. Given a kernel function $k(x, y)$ that satisfies the properties of a valid kernel, the kernel matrix K is constructed such that each entry K_{ij} is $K_{ij} = k(x_i, x_j)$, for vectors x_i and x_j in the input space. Which of the following properties is always true for the kernel matrix K?

- ☐ The smallest eigenvalue is always zero.
- ■ The matrix is symmetric.
- ☐ There is at least one negative eigenvalue.
- ■ The matrix is invertible.

18. Consider a dataset with 10 instances, 6 of which belong to Class A and 4 belong to Class B. You are evaluating a potential split for a decision tree. This split divides the dataset into two groups: one with 2 instances of Class A and 2 of Class B, and another with 4 instances of Class A and 2 of Class B. Calculate the information gain from this split using entropy as the impurity measure. What is the information gain of the split?

- ☐ $-\left(\frac{5}{10}\log_2\left(\frac{5}{10}\right)+\frac{5}{10}\log_2\left(\frac{5}{10}\right)\right)$ $-\left(\frac{4}{10}\left(-\frac{2}{4}\log_2\left(\frac{2}{4}\right)-\frac{2}{4}\log_2\left(\frac{2}{4}\right)\right)\right)$ $-\frac{6}{10}\left(-\frac{3}{6}\log_2\left(\frac{3}{6}\right)-\frac{3}{6}\log_2\left(\frac{3}{6}\right)\right)$
- ☐ $-\left(\frac{4}{10}\log_2\left(\frac{4}{10}\right)+\frac{6}{10}\log_2\left(\frac{6}{10}\right)\right)$ $-\left(\frac{4}{10}\left(-\frac{3}{4}\log_2\left(\frac{3}{4}\right)-\frac{1}{4}\log_2\left(\frac{1}{4}\right)\right)\right)$ $-\frac{6}{10}\left(-\frac{2}{6}\log_2\left(\frac{2}{6}\right)-\frac{4}{6}\log_2\left(\frac{4}{6}\right)\right)$
- ■ $-\left(\frac{6}{10}\log_2\left(\frac{6}{10}\right)+\frac{4}{10}\log_2\left(\frac{4}{10}\right)\right)$ $-\left(\frac{4}{10}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right)-\frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\right)$ $-\frac{6}{10}\left(-\frac{4}{6}\log_2\left(\frac{4}{6}\right)-\frac{2}{6}\log_2\left(\frac{2}{6}\right)\right)$
- ☐ $-\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right)+\frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)$ $-\left(\frac{4}{10}\left(-\frac{1}{4}\log_2\left(\frac{1}{4}\right)-\frac{3}{4}\log_2\left(\frac{3}{4}\right)\right)\right)$ $-\frac{6}{10}\left(-\frac{1}{6}\log_2\left(\frac{1}{6}\right)-\frac{5}{6}\log_2\left(\frac{5}{6}\right)\right)$



19. Consider the properties of ROC (Receiver Operating Characteristic) curves for binary classifiers.

- ■ The ROC curve of a binary classifier is better than random guessing when the area under the curve (AUC) is above 0.5.
- ☐ The ideal point on a ROC curve is at the top right corner, representing a classifier with perfect recall and no false positives.
- ☐ A ROC curve that bows towards the top left corner indicates a classifier with perfect precision, but not necessarily perfect recall.
- ☐ The area under the ROC curve of a classifier that performs at chance level is 0.5.

20. Imagine we are fitting a linear regression model and choose not to assume equal significance for all weight vectors. Instead, we implement a method that imposes a Gaussian-distributed prior to influence the estimation of the regression coefficients. Which two of the following methods are we likely using?

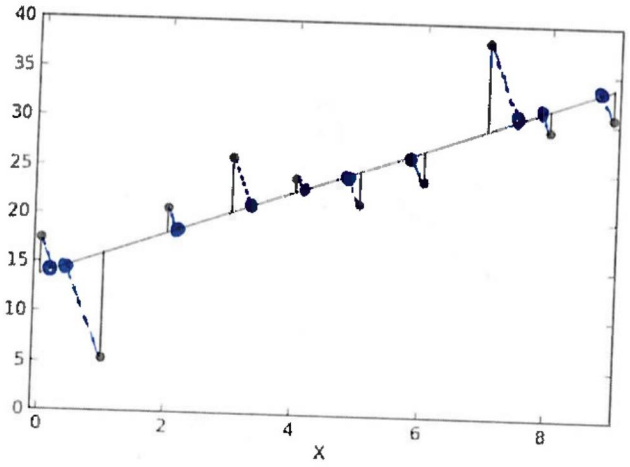
- ☐ Elastic Net regularization.
- ■ Ridge regression.
- ■ Bayesian regression with normally-distributed priors.
- ☐ Weighted least squares.

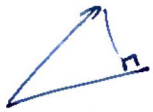
2. **Linear Least Squares vs. PCA.** (3 + 6 + 4 = 13 points) Consider a dataset consisting of n points (x_i, y_i) with coordinates $x_i \in \mathbb{R}$ and $y_i \in \mathbb{R}$.

(a) Given y as the response variable, provide the formulation for least squares regression to model the relationship between x and y. State the objective function that the linear least squares method optimizes.

(b) Describe the mathematical objective function that PCA optimizes when performing dimensionality reduction from $\mathbb{R}^2$ to $\mathbb{R}$. Outline the steps for performing PCA on this dataset.

(c) Given the picture for linear least squares, sketch the corresponding illustration for PCA in (b).





a) add a bias term (a 1) to each feature vector, model.

$$y = \beta^T X + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

Objective: minimize the RSS

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2 \quad \text{in matrix form} \quad \underset{\beta}{\operatorname{argmin}} (y - X\beta)^T (y - X\beta)$$

b) Let X be the data matrix, $\Sigma = \frac{1}{n} X^T X$ be the sample covariance matrix. PCA aims to find a vector w s.t

$$\hat{w} = \underset{w}{\operatorname{argmax}}\ w^T \Sigma w \leftarrow \text{variance.} \qquad (*)$$

If there are already n vectors $\hat{w}_1, \hat{w}_2 \ldots \hat{w}_n$ determined PCA adds an additional constraint that $\hat{w}$ needs to be orthogonal to $\hat{w}_1, \hat{w}_2 \ldots \hat{w}_n$

Since in the problem, PCA aims to reduce $\mathbb{R}^2$ data to $\mathbb{R}$. we need to first compute the covariance matrix Σ then solve for

$$\hat{w} = \underset{w}{\operatorname{argmax}}\ w^T \Sigma w$$

← no need to impose orthogonality since this is first vector

which can be done w/ SVD or eigendecomposition

Additional space for Problem 2.

c) assuming the line represents the direction for which the projected points achieves the maximum variance, then the PCA result is given by the projection of the points onto the line (see the graph)

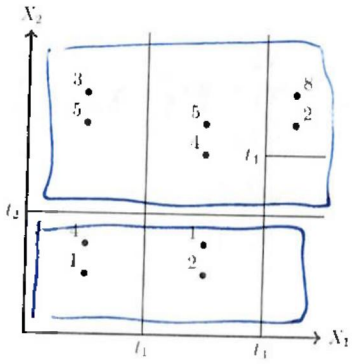
If the line is not the principal component: need to first solve

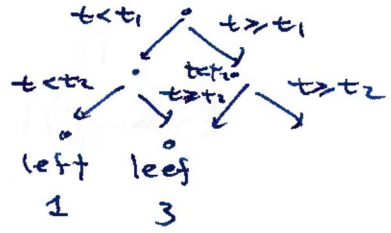
$$\hat{w} = \arg\max_{w} w^T \frac{X^T X}{n} w.$$

which to get the direction $\hat{w}$ → using SVD or eigendecomposition

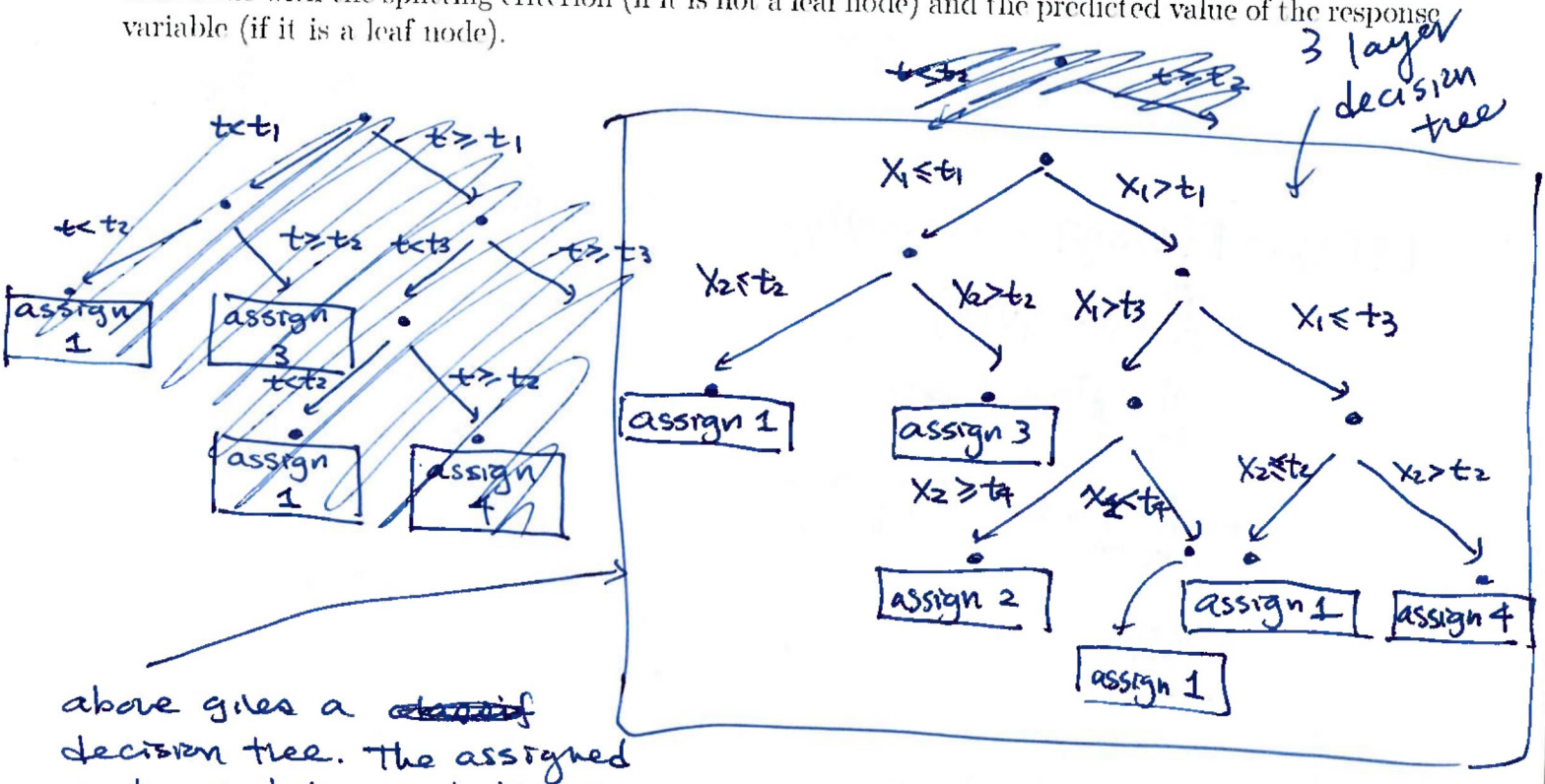
$\frac{10}{140}$ $\frac{10}{14}$ 0.7 5/7 71%

3. **Decision Tree Construction from Feature Space.** (7 points) Below is a representation of a two-dimensional feature space for a dataset, divided into partitions. The features X_1 and X_2 correspond to the horizontal and vertical axes, respectively. Each dot represents a data point with a corresponding response variable.



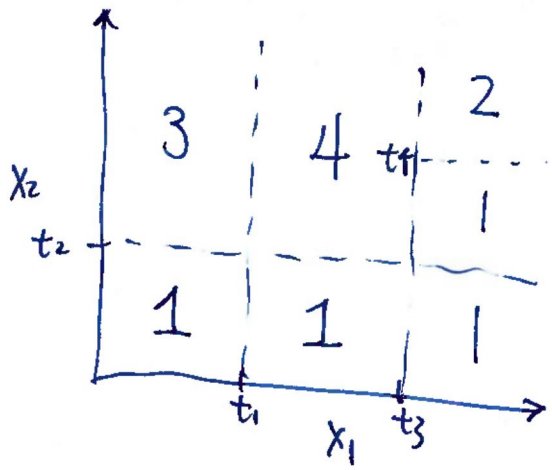


Draw the simplest decision tree that represents the partitioning shown in the feature space. Label each node with the splitting criterion (if it is not a leaf node) and the predicted value of the response variable (if it is a leaf node).



above gives a ~~classif~~ decision tree. The assigned value is determined by choosing the most frequent class in each region. (If tie, a random one is chosen)

The predicted value by region is shown here →



There are 5 errors made, so classification error is 50%

4. Linear Regression with Laplacian-Distributed Errors. (4 + 5 points)

Question: Consider the linear regression model $y_i = \mathbf{x}_i^T \beta + \epsilon_i$ applied to the dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^p$ is a feature vector, $y_i \in \mathbb{R}$ is the observed outcome, and ϵ_i are independent and identically distributed (i.i.d.) errors. Assume that the errors follow a Laplacian distribution, with the probability density function defined as:

$$f(\epsilon_i; b) = \frac{1}{2b} \exp\left(-\frac{|\epsilon_i|}{b}\right),$$

where $b > 0$ is the scale parameter.

(a) Derive the likelihood function $L(\beta; \mathcal{D})$ for the parameter vector β assuming a Laplacian distribution for the errors. Provide the formal expression for this likelihood.

(b) Formulate the simplest possible objective function that must be minimized to estimate the regression parameters β under the Laplacian error model. Is there an explicit formula for the minimizer ? Compare with ordinary least squares.

a) 1. ~~$\mathbb{P}\{X = x\} \Rightarrow \mathbb{P}\{X \le a\} = \mathbb{P}\{X \le a\} = F_X(a)$~~

$f_X(a) = f_X(a+\epsilon)$

~~this implies that $\beta_i = \beta_j$~~

$$\mathcal{L}(\beta, D) = \prod_{i=1}^{n} \mathbb{P}\{D_i \mid \beta\} = \prod_{i=1}^{n} \frac{1}{2b} \exp\left(-\frac{|y_i - x_i^T \beta|}{b}\right)$$

(D_i: ith dataset)

$$= \left(\frac{1}{2b}\right)^n \exp\left(-\frac{1}{b}\sum_{i=1}^{n} |y_i - x_i^T \beta|\right)$$

The problem thus reduces to

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} |y_i - x_i^T \beta|.$$

b) The regression parameter $\hat{\beta}_L$ solves.

$$\hat{\beta}_L = \arg\max_{\beta} \mathcal{L}(\beta, D)$$

$$= \arg\max_{\beta} \left(\frac{1}{2b}\right)^n \exp\left(-\frac{1}{b}\sum_{i=1}^{n} |y_i - x_i^T \beta|\right)$$

Consider log-likelihood.

$$\ell(\beta, D) = \log(\mathcal{L}(\beta, D)) = n \log\left(\frac{1}{2b}\right) - \frac{1}{b}\sum_{i=1}^{n} |y_i - x_i^T \beta|$$

This maximization does not have an explicit solution since the log-likelihood involves an absolute value, which is not always differentiable.

However, iterative method can be used to estimate the optimal parameter.

Additional space for Problem 4.

In ordinary least squares the $\hat{\beta}$ is obtained by minimized

$$\hat{\beta} = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

In this scenario, $\hat{\beta}$ is obtained by minimizing

$$\hat{\beta} = \sum_{i=1}^{n} | y_i - x_i^T \beta |$$

Both method is minimizing

$$\hat{\beta} = \sum_{i=1}^{n} l(y_i, x_i^T \beta)$$

But the loss function is just different ~~for both~~, one uses square loss and another uses absolote value loss.

In some sense, this is the "average" loss if multiply by $\frac{1}{n}$. ~~So~~

i.e. $\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} l(y_i, x_i^T \beta)$

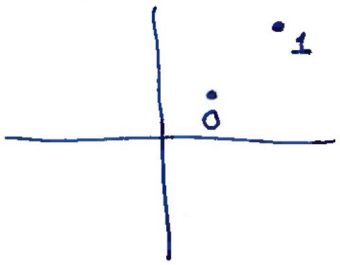
So $\hat{\beta}$ is the minimizer of the average loss.

5. **VC dimension.** (8 points) Let $\mathcal{F}$ be the set of classifiers corresponding to all concentric circles in the plane (that is, $\mathbb{R}^2$) centered at the origin, precisely

$$f_r(x) := \begin{cases} 1, \|x\|_2 \leq r \\ 0, \text{otherwise.} \end{cases}$$

Find the VC dimension of this set of classifiers.

VC dimension is 1 since it cannot shatter 2 points, consider the example below



any classifier of type f_r will misclassify one of the data point ~~by response~~ .0

6. **Margin Dynamics in Boosting Algorithms.** (5 + 4 = 9 points) In the context of boosting algorithms, the margin of a classification can be described as the extent to which combined weak classifiers prediction for a data point is correct and confident. Consider a boosting algorithm applied to a binary classification task with labels $y_i \in \{-1, +1\}$ and a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$.

(a) Define the concept of margin in the context of boosting classifiers. How is the margin for a data point $(\mathbf{x}_i, y_i)$ mathematically expressed in terms of the weak classifiers' weighted votes? Provide a detailed explanation.

(b) Explain why a boosting algorithm might achieve a low testing error quickly compared to training error, in terms of the margins provided by the weak classifiers.

a) Let $G_1, G_2 \ldots G_T$ be weak classifiers, in boosting, we consider their weighted vote

$$G(x) = \text{sign}\left(\sum_{t=1}^{I} \alpha_t G_t(x)\right)$$

A sample (x_i, y_i) is correctly classified if $y_i G(x_i) = \pm 1$

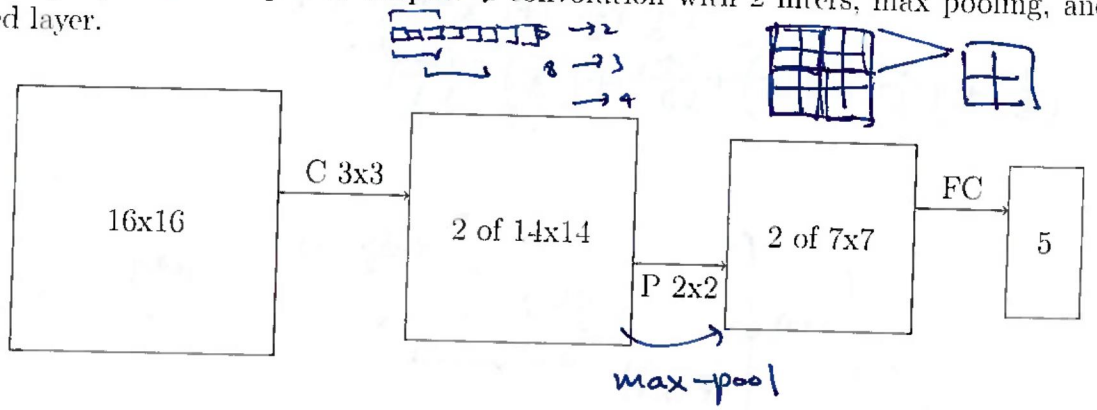
The empirical # of missclassifications is given by

$$\sum_{i=1}^{n} \mathbb{1}\{y_i G(x_i) \neq 1\}$$

Let N denote the # of misstclassification

b) when the overall margin is high, this means that the combined model (using the weak learners) are confident w/ the correct classification. This means that the test rate drops significantly, much faster compared to the traing error.

7. **Convolutional Neural Network.** (2+2+2+2+2 = 10 points) Consider a simplified convolutional neural network that processes a 16x16 image into a set of class scores. The network comprises the following layers from input to output: a convolution with 2 filters, max pooling, and a fully-connected layer.



The convolutional layer applies 2 filters of size 3x3 to the input image and creates two 14 by 14 images. Each filter moves across the image one pixel at a time (a stride of 1), which means every pixel is covered by the filter for the convolution operation. After the convolution, a max pooling operation with a 2x2 window is applied (with a stride of 2). Finally, a fully-connected layer produces the output class scores.

(a) Calculate the total number of weights in the convolutional layer.

(b) Explain how changing the stride of the convolutional layer from 1 to 2 would affect the output dimensionality of the convolutional layer.

(c) Determine the number of max pooling operations performed in the max pooling layer during the forward pass.

(d) Compute the number of parameters in the fully connected layer.

(e) If the dropout method with parameter $p = 1/2$ is used as a regularization method, how will the total number of parameters in the trained network change?

a) In the convolution layer, we only need to train the filter of size 3x3, so 9 parameters assuming no bias layers. for weight.

If bias is set to true, then there will be 18 parameters (a 3X3 filter and a 3x3 bias)

b) The dimensionality would be lowered. since after one region has been applied the filter, the filter moves 2 steps to the right. This essentially reduces the number of outputted image (ex: if the image size is 8X8, stride 1 gives dimension of. image size of 6x6, stride 2 gives 3X3)

c) For each 14x14 image, there are 49 max pool operations, so there are 98 max pool operations for 2 images.

Additional space for Problem 7.

d) ~~The~~
~~After flattening the image, there are~~.

This depends on the specific architecture design of the network.

Assuming that we flatten each of the 2 images and concatenate the 2 49×1 dimension vector, we have input shape of 98×1.

Assuming there are 5 neurons, then each neuron contains 98 weight parameters and 1 bias parameter.

So in ~~total~~, there are $\underbrace{98 \times 5}_{\text{weight parameters}} + \underbrace{1 \times 5}_{\text{bias parameter}} = \underbrace{99 \times 5}_{\text{total}}$ ~~for~~ parameters to train.

e) The total # of parameters will remain the same, however during training, ~~some~~ in each round, some parameters may be "dropped" with some probability to prevent overfitting.

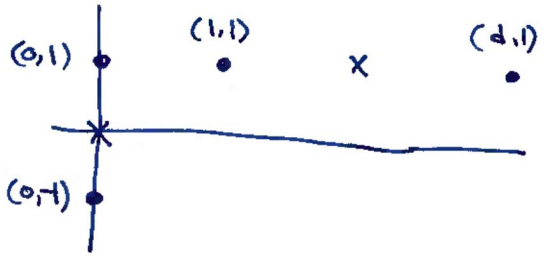
8. **Convergence Properties of Lloyd's Algorithm.** (6 + 6 = 12 points)

(a) Demonstrate with a formal example that Lloyd's k-means clustering algorithm may converge to a set of cluster centroids that do not correspond to the global minimum of the objective function.

(b) Is it possible for Lloyd's algorithm to fail to converge to a local minimum? Provide a formal example, including relevant computations, to illustrate your answer.

In both cases, provide formal arguments supporting your claims.

a) casider 2-means clustering, casider the follong configuration



Assme the algorithm starts w/ centroid (0,1), (1,1), then

~~→ (0,1) casidore~~

→ group 1: (0,1), (0,-1)

→ group 2: (1,1), (d,1)

Now computag the centroid gues.

→ group 1: (0,0)

→ group 2: $(\frac{d+1}{2}, 0)$

For $\frac{d+1}{2} - 1 \leq \sqrt{2}$, Lloyd's algorithm converges

equivalently $d \leq 1+2\sqrt{2}$

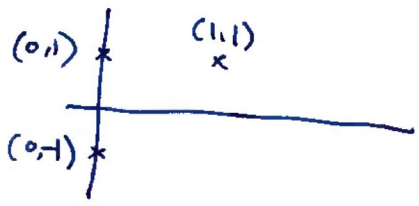
Continue next page

b) No, Lloyd's algorithm always converge to the local minimum. As shown in HW5, dung traing, the objective — the inter-cluster variance, is strictly decreasing each iteration, so it must converge to a local minimum (it may not converge to the global minimum through, so it's nessesary to run the algorithm multiple times).

sum of

Additional space for Problem 8.

However, when $2(d-1)^2 \geq \frac{1^2+2^2+1^2+4^2+2^2+2^2}{3^2}$ or $d \geq 1+\sqrt{\frac{2}{3}}$,



it can be computed that the cluster

Group 1: (-,1), (1,1), (0,-1)
Group 2: (d,1)

minimizes the intercluster variance.

Since $1+\sqrt{\frac{2}{3}} \leq d \leq 1+2\sqrt{2}$ is non-empty, as long as we choose $d \in (1+\sqrt{\frac{2}{3}}, 1+2\sqrt{2})$, we see that

① since $d \leq 1+2\sqrt{2}$, the converged solution using Lloyd's algorithm is

Group 1: (0,1), (0,-1) w/ centroid (0,0)
Group 2 (1,1), (d,1) w/ centroid $(\frac{d+1}{2}, 1)$

② since $d \geq 1+\sqrt{\frac{2}{3}}$, the optimal solution is given by

Group 1: (0,1), (1,1) (0,-1) w/ centroid $(\frac{1}{3}, \frac{1}{3})$
Group 2: (d,1) w/ centroid (d,1)

→ achieves lower intercluster variance.

This ~~suggest~~ gives a numerical example for which Lloyd's algorithm fails to converge to global optimal.