Recitation 13- HW7 and Final Exam Review

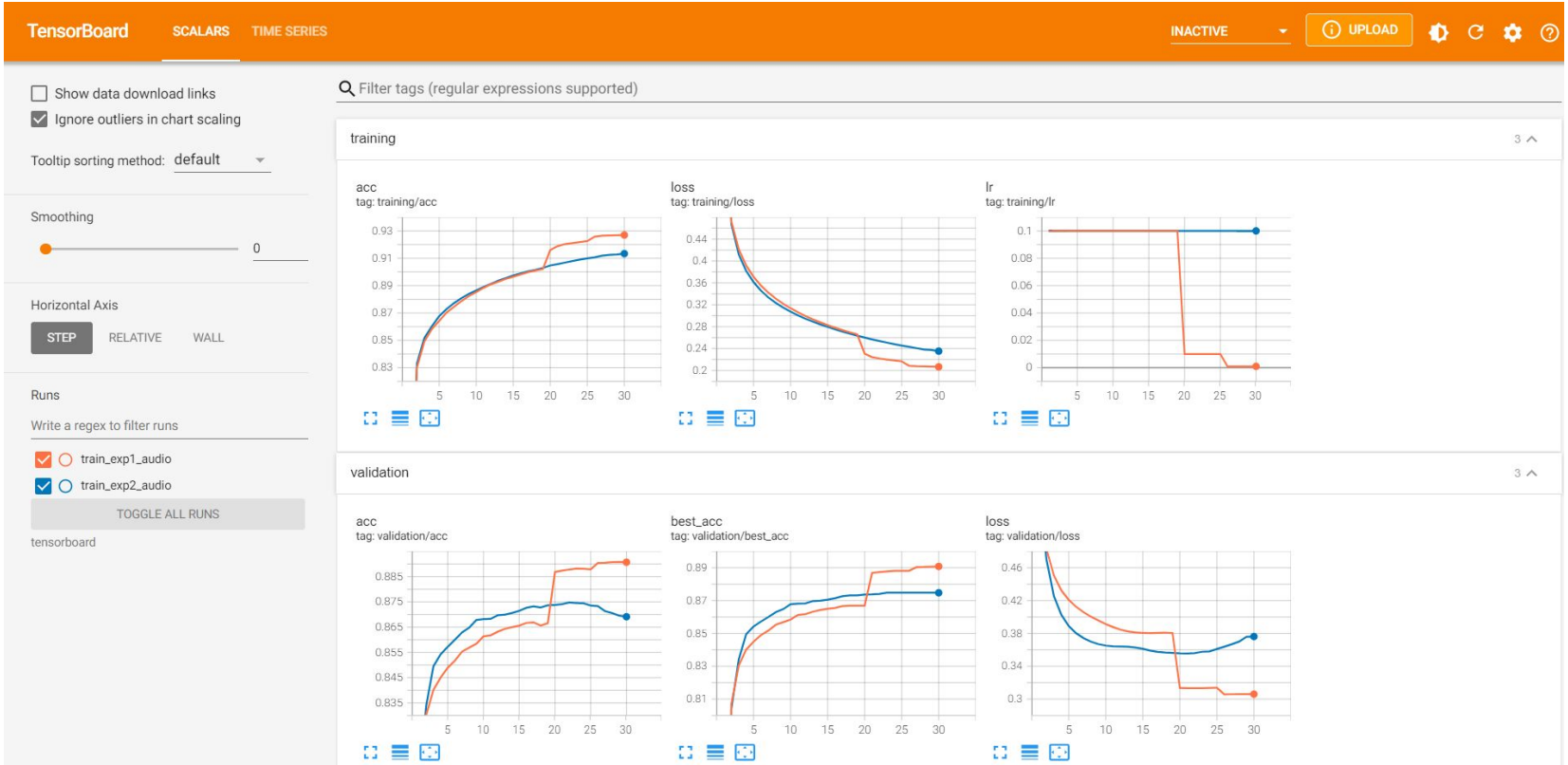
Roshan Sharma, Ching-yi Lin, Courage Ekoh

HW-7 Review

- Problem 1: MLP Classifier with Pytorch
  BONUS 1.5- CNN Classifier with Pytorch
- Problem 2: Improving on MLP Classification with Hyperparameter Tuning
- Problem 3: Decision Trees with scikit-learn on Spotify data
- Problem 4: K-Means clustering with COVID-19 data

You must write and submit **your own code**

Tensorboard Visualizations



Questions on HW-7 ?

Exam Review

Part 1: MLE/MAP, Naive Bayes, Logistic Regression, Linear regression, Bias/variance tradeoff

Summary of Methods

MLE:

- Given a set of data points, we first select a distribution that fits the data , and find the parameters by maximizing the likelihood of the data points
- $L(\Theta) = P(D \mid \Theta) = P(x_1 \mid \Theta) P(x_2 \mid \Theta) .....P(x_N \mid \Theta)$- independence assumption

MAP:

- Maximize posterior $P(\Theta|X) = L(\Theta) P(\Theta)$
- Prior information is encoded in $P(\Theta)$
- Prior represents expert knowledge (philosophical approach) or is chosen to ensure a simple posterior form (engineer's approach)

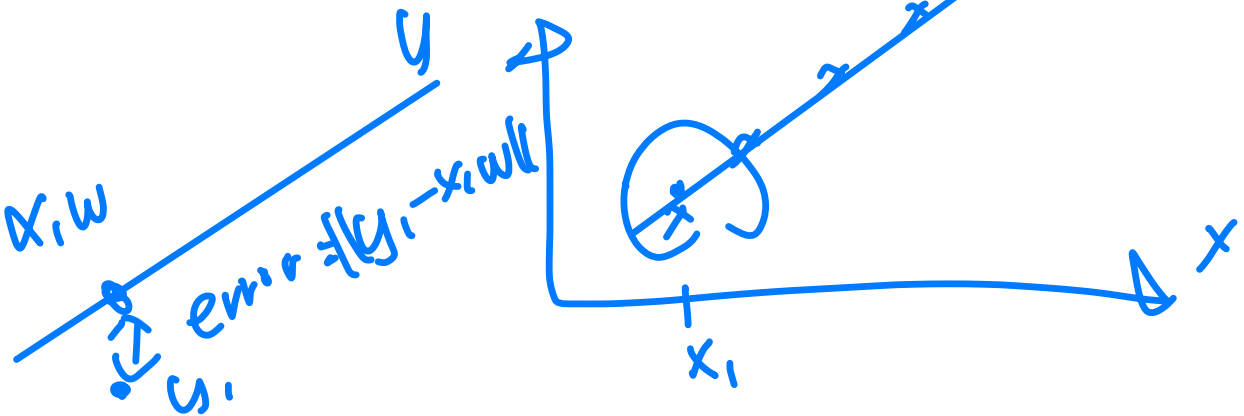
Summary of Methods

Linear Regression:

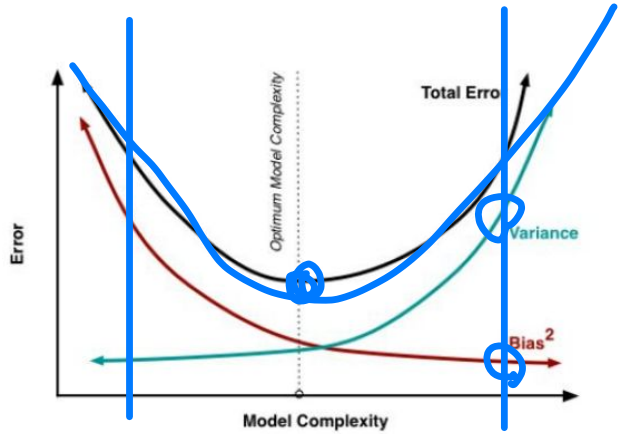
Find the parameters w that minimize the sum of least squares $|| y- Xw||^2_2$

Regularization by adding a penalty on weights- RLS: $|| y- Xw||^2_2 + k ||w||^2_2$

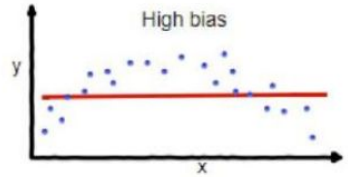
Linear Regression can have probabilistic interpretations



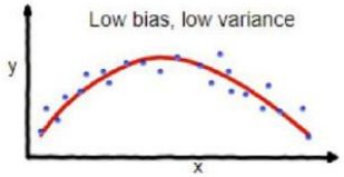
Summary of Methods



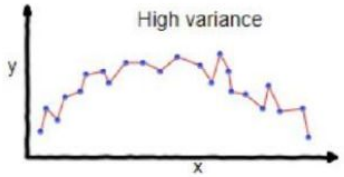
Bias- Variance Tradeoff



underfitting



Good balance



overfitting

Comparison of Methods

Logistic Regression:

- Classification method that uses the logistic loss
- Discriminative Approach
- Makes no assumptions and generally lower asymptotic error
- Slower convergence
- Linear Decision Boundary

$$P(Y = 1|X) = \frac{1}{1 + \exp(w_0 + \sum_i w_i X_i)}$$

Naive Bayes:

- Classification method that assumes conditional independence among feature components
- Generative approach
- Naïve Bayes makes more restrictive assumptions and has higher asymptotic error
- But it converges faster
- Linear Decision Boundary

$$P(Y = y_k|X_1 \dots X_n) = \frac{P(Y = y_k) \prod_i P(X_i|Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i|Y = y_j)}$$

MCQ-1

$ax^2+bx+c=0 \Rightarrow x = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$

$A = (X^T X) X y$

Which of the following statements is TRUE?

- ○ We can find a closed-form solution to the maximum-likelihood estimator of a logistic regression model.
- ○ The objective in training the logistic regression model is to find the parameters $\theta^* = \text{argmax}_\theta J(\theta)$, where $J(\theta)$ is the negative logarithm of the cross-entropy loss function. argmin
- ✓ The output of the logistic regression model can provide a probabilistic interpretation of the predicted classification label.
- ○ The binary logistic regression model (without kernels) is a nonlinear classifier, i.e., the boundary between two classes is not a linear one.

$\frac{1}{1+\exp(\Sigma wx)} > .5 \Leftrightarrow \exp(\Sigma wx) > 1 \Leftrightarrow \Sigma wx > 0$

MCQ-2

Assume we have a coin that lands as "heads" with probability θ when flipped. To estimate the value of θ, we independently flip this coin n times, which yields the observation set $\mathcal{D}$ with n_H heads and n_T tails. We also have prior knowledge that θ follows the Beta distribution $B(\theta; \alpha, \beta)$, which has the probability density function $f(\theta; \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$. Which of the following expressions is proportional to the resulting posterior distribution $P(\theta|\mathcal{D})$?

- $\theta^{n_H}(1-\theta)^{n_T}$
- $\theta^{n_T}(1-\theta)^{n_H}$
- $\theta^{\alpha+n_H-1}(1-\theta)^{\beta+n_T-1}$
- $\theta^{\beta+n_T-1}(1-\theta)^{\alpha+n_H-1}$

$\alpha = \beta = 1000000$

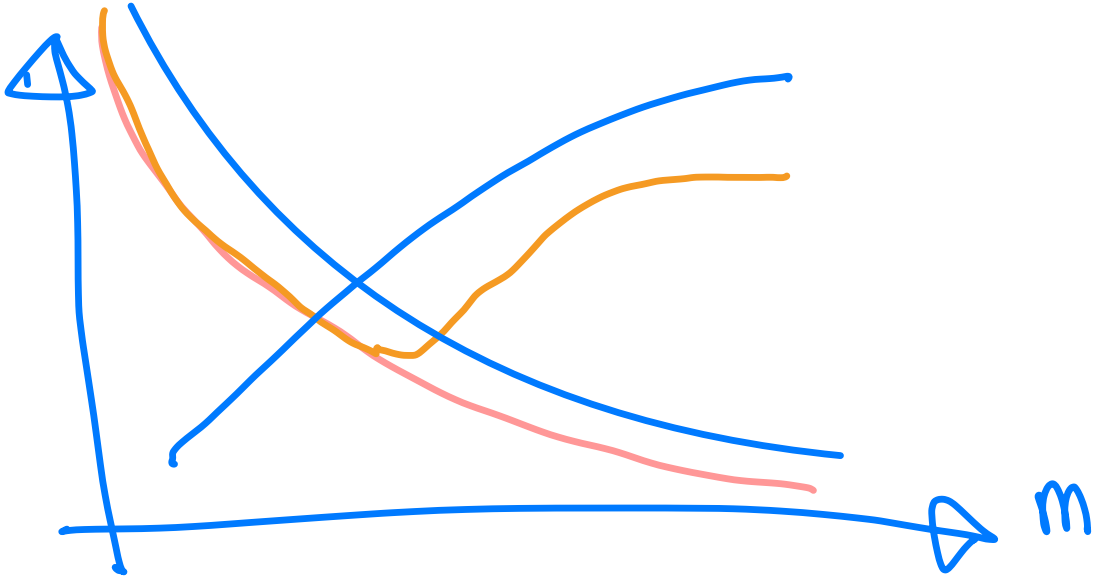
$$\mathcal{L}(\theta) = P(\mathcal{D}|\theta)$$
$$= P(x_1|\theta) \cdot P(x_2|\theta) \cdots P(x_N|\theta)$$
$$= \theta^{n_H}(1-\theta)^{n_T} \quad \theta^{\alpha-1}(1-\theta)^{\beta-1}$$
$$= \theta^{n_H+\alpha-1}(1-\theta)^{n_T+\beta-1}$$

$\theta^{n_H}(1-\theta)^{n_T}$

MCQ-3

Suppose you are performing least-squares polynomial regression, i.e., you are estimating the parameters $\beta_0, \beta_1, \ldots, \beta_m$ that best fit the model $y = \sum_{i=0}^{m} \beta_i x^i$ for a given set of (x, y) pairs. As the degree m of your polynomials increases, which of the following quantities would you expect to initially decrease but then increase?

- ○ Training error
- ○ Validation error
- ○ Bias
- ○ Variance



MCQ- 4

Problem 4: What is the formula for the logistic regression decision boundary in the x_1, x_2 plane for:

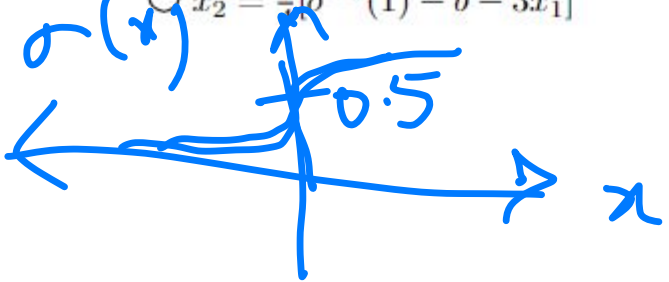
$$\Pr(y = 1 \mid x_1, x_2) = \sigma(b + 3x_1 + 4x_2)$$

Note that σ^{-1} is the inverse of the sigmoid function.

- $x_2 = \frac{1}{4}[\sigma^{-1}(0.5) - b - 3x_1]$
- $x_2 = \frac{1}{4}[\sigma^{-1}(0.75) - b - 3x_1]$
- $x_2 = \frac{1}{4}[\sigma^{-1}(0.25) - b - 3x_1]$
- $x_2 = \frac{1}{4}[\sigma^{-1}(1) - b - 3x_1]$

$w_1 = 3$

$w_2 = 4$



$4x_2 + 3x_1 + b = 0$

$\sigma \rightarrow \sigma^{-1}(0.5) = 4x_2 + 3x_1 + b$

$x_2 = \frac{1}{4}[\sigma^{-1}(0.5) - b - 3x_1]$

MCQ- 4

Problem 4: What is the formula for the logistic regression decision boundary in the x_1, x_2 plane for:

$$\Pr(y = 1 \mid x_1, x_2) = \sigma(b + 3x_1 + 4x_2) = \frac{1}{2}$$

Note that σ^{-1} is the inverse of the sigmoid function.

- ✓ $x_2 = \frac{1}{4}[\sigma^{-1}(0.5) - b - 3x_1]$
- $x_2 = \frac{1}{4}[\sigma^{-1}(0.75) - b - 3x_1]$
- $x_2 = \frac{1}{4}[\sigma^{-1}(0.25) - b - 3x_1]$
- $x_2 = \frac{1}{4}[\sigma^{-1}(1) - b - 3x_1]$

$\sigma(b + 3x_1 + 4x_2) = 0.5$

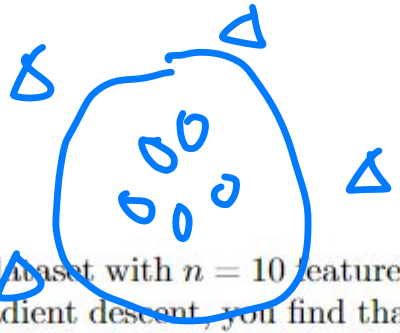
$\Rightarrow b + 3x_1 + 4x_2 = \sigma^{-1}(0.5) = 0$

MCQ-5

Which of the following statements about Naive Bayes is FALSE ?

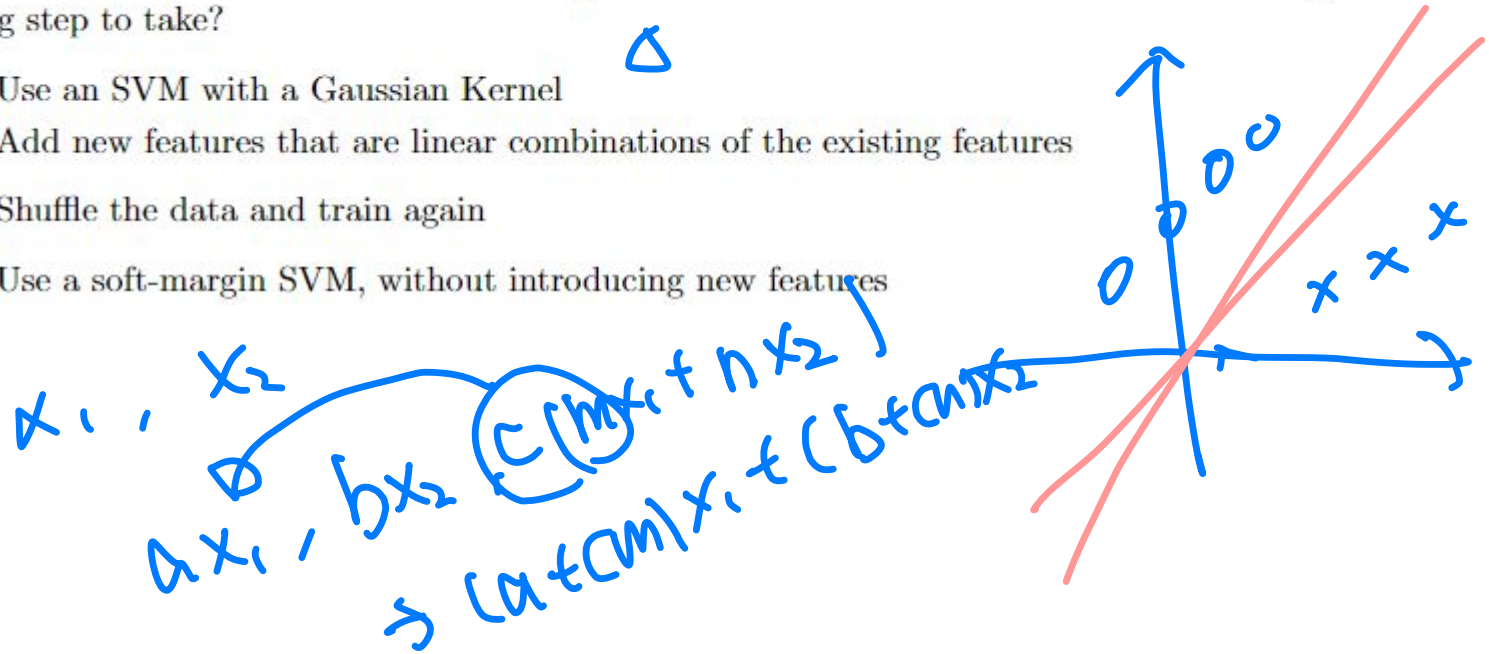
1. We do not have any prior knowledge (excluding the data itself) on whether any feature is more important than the others. (O)
2. The features are assumed to be statistically independent of one another given the class label. (O)
3. Suppose a feature appears multiple times within one data sample. Naive Bayes accounts for statistical dependencies between these multiple occurrences. (X)
4. Naive Bayes is a probabilistic model (O)

MCQ-6



Problem 10: Suppose you have a dataset with $n = 10$ features and $m = 5000$ examples. After training your logistic regression classifier with gradient descent, you find that it has underfit the training set and does not achieve the desired performance on the training or cross validation sets. Which of the following is the most promising step to take?

- ○ Use an SVM with a Gaussian Kernel
- ○ Add new features that are linear combinations of the existing features
- ○ Shuffle the data and train again
- ○ Use a soft-margin SVM, without introducing new features

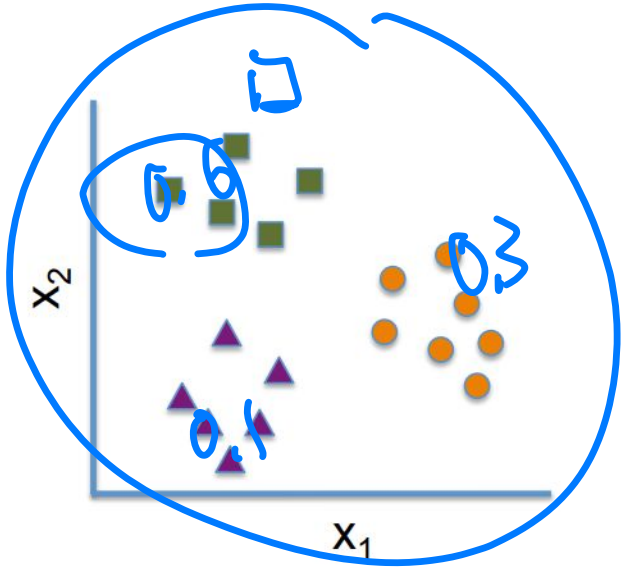


Part 2: Multiclass, SVM, Graphical models, Nearest neighbors, Decision trees, Boosting/ensemble methods

Multi-class Classification

Approaches

- One-versus-one
- One-versus-all



MCQ-1

Problem 4: *[4 points]* In lecture, we studied methods of extending Naïve Bayes and logistic regression models to multi-class ones that can predict a data point's membership in one of multiple (i.e., more than two) possible classes. One such method is called *multinomial logistic regression* (also known as softmax regression). The two other extension methods for logistic regression, *one-vs.-one* and *one-vs.-all*, decompose the multi-class classification problem into multiple binary classification ones.

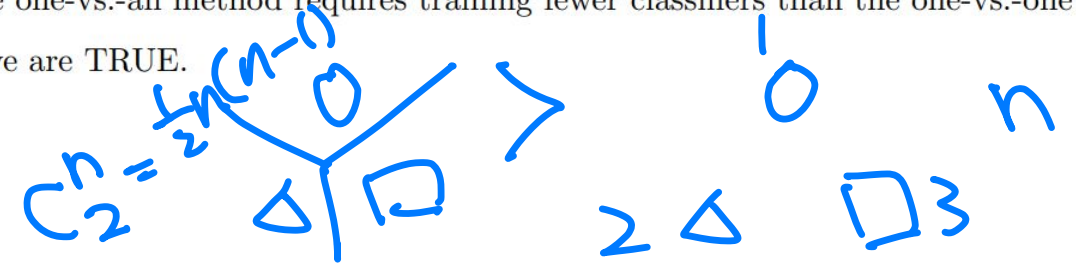
Consider a classification problem with $C \geq 2$ classes. Which of the following statements about these multi-class classification techniques is TRUE?

a. Both multinomial logistic regression and Naïve Bayes output the probabilities that a point x belongs to each class $c = 1, 2, \ldots, C$.

b. If $C = 2$, multinomial logistic regression and the one-vs.-one classification method could give different solutions.

c. When $C > 2$, the one-vs.-all method requires training fewer classifiers than the one-vs.-one method.

d. None of the above are TRUE.



MCQ-2

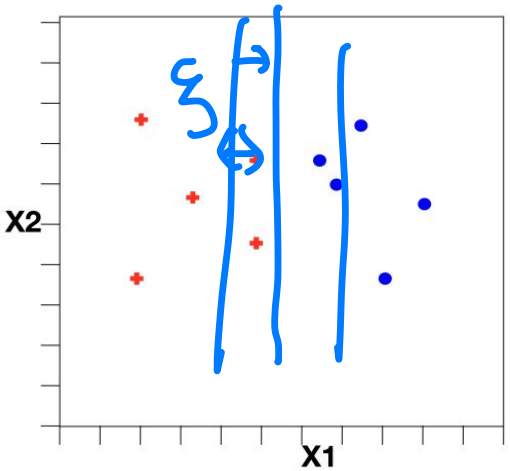
Consider the general optimization problem of soft-margin SVM:

$$\min_{\mathbf{w},b,\xi_i} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i$$

$$\text{subject to } y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \quad \forall i$$

$$\xi_i \geq 0 \quad \forall i,$$

What happens to the width of the margin as $C \to 0$?



MCQ-3

Problem 3: *[8 points]* You randomly interviewed 10 CMU students and obtained the following statistics. Based on a student's course schedule, average sleep hours, internship/full time offer status, and number of finals, we want to predict the student's mood with a decision tree.

Taking 18661?	Avg sleep hrs	Offer status	#finals	Mood
Yes	>7	Has Offer	≥2	Happy
No	≤7	No Offer	≥2	Sad
Yes	≤7	Has Offer	<2	Sad
No	>7	No Offer	≥2	Sad
Yes	>7	No Offer	<2	Happy
Yes	>7	Has Offer	<2	Happy
No	≤7	No Offer	<2	Sad
No	>7	Has Offer	<2	Happy
Yes	>7	No Offer	≥2	Sad
No	≤7	Has Offer	≥2	Happy

a. *[2 points]* What is the initial entropy of mood, using base-2 logarithms?

b. *[3 points]* Find the information gain (IG) from splitting on "offer status". You may leave $\log_2$ expressions in your answer.

c. *[3 points]* Using "offer status" as the root of the tree, can you perfectly predict a student's mood using a decision tree with 2 splits? Briefly justify why or why not. You do not need to calculate information gains in your answer.

Solution:

a. We solve for entropy for our label with: $H(y) = -\sum_{i \in [y]} P(y = i) \log_2 P(y = i)$

$P(\text{happy}) = \frac{1}{2}, P(\text{sad}) = \frac{1}{2}$

The initial entropy of mood (H(mood)) is $-\frac{1}{2}\log_2 \frac{1}{2} - \frac{1}{2}\log_2 \frac{1}{2} = 1$.

b. The conditional probability of $H(Y|X)$ for each attribute label k is calculated by:

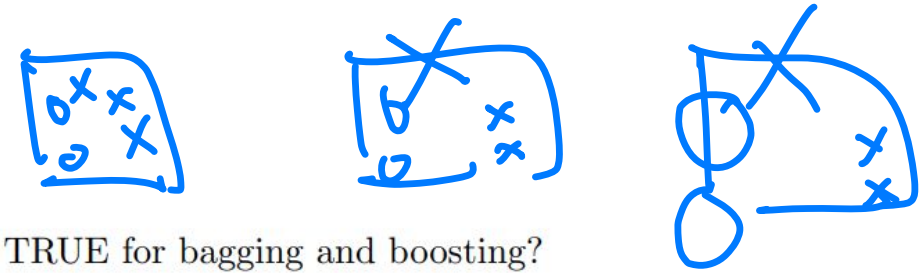
$H(Y|X) = \sum_k P(X = a_k) H(Y|X = a_k)$

$\text{H(mood|offer status)} = -\frac{1}{2}\left(\frac{1}{5}\log\frac{1}{5} + \frac{4}{5}\log\frac{4}{5}\right) - \frac{1}{2}\left(\frac{1}{5}\log\frac{1}{5} + \frac{4}{5}\log\frac{4}{5}\right)$

Information gain is $\text{I(Usage}, X_i) = \text{H(Usage)} - \text{H(Usage}|X_i) = 1 + \frac{1}{2}\left(\frac{1}{5}\log\frac{1}{5} + \frac{4}{5}\log\frac{4}{5}\right) + \frac{1}{2}\left(\frac{1}{5}\log\frac{1}{5} + \frac{4}{5}\log\frac{4}{5}\right)$.

c. No we can't. After splitting root with offer status, we still cannot fully categorize a student's mood with any one of the remaining three criteria.

MCQ-4



Which of the following statement is TRUE for bagging and boosting?

- Random Forest creates different boostraps by sampling from the training set without replacement and train each regression tree with a random subset of features.
- AdaBoost trains weak classifiers in parallel, each weak classifier attempts to minimize the predicton error of weighted input.
- If each classifier is (at least) slightly better than random, AdaBoost will achieve zero training loss eventually.
- Bagging aims to decrease the bias of the original classifer.

Part 3: Neural networks, Clustering, EM, Dimensionality reduction, Reinforcement, online learning

Summary of Methods

Neural Network:

- A supervised learning method for classification or regression which mimics human brain structure
- Multi-layer perceptron comprise perceptrons with non-linear activations

PCA:

- A dimensionality reduction method
- Reduces data to the dimension with the largest variance

Summary of Methods

Reinforcement Learning:

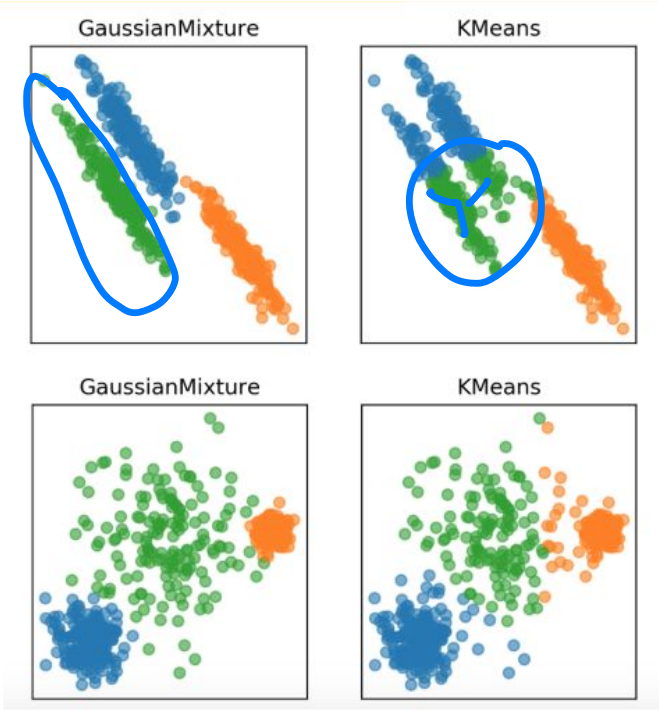
- Different with supervised and unsupervised learning
- Based on rewarding desired behaviors and/or punishing undesired ones

EM:

- Splits an optimization into E-step and M-step
- Iteratively maximizes each step guarantee improvement

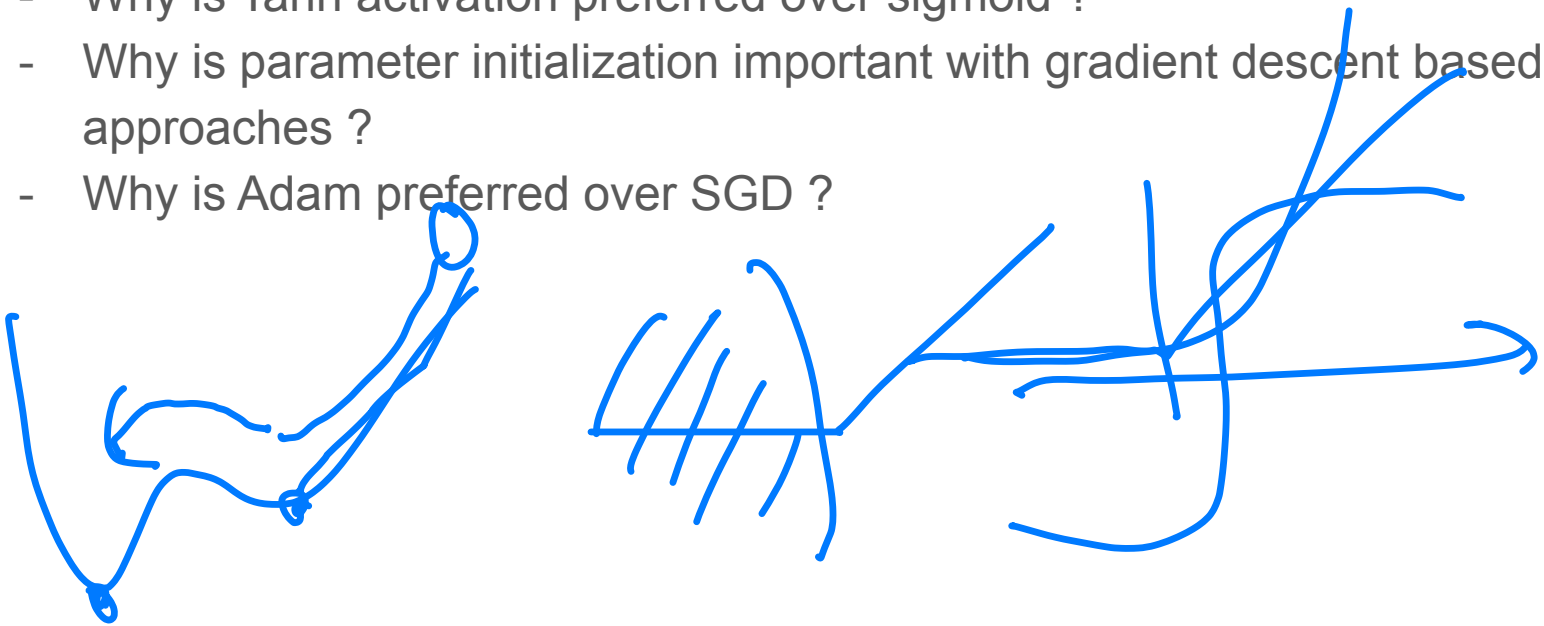
Comparison between K-Means and GMM

	K-Means	GMM
Candidates	Closer	Gaussian prob
Cluster characteristics	Mean	Mean, variance
Handle scaled dimension	X	V



Conceptual Question: Neural Networks

- Why is ReLU activation preferred over Sigmoid and Tanh activations ?
- Why is Tanh activation preferred over sigmoid ?
- Why is parameter initialization important with gradient descent based approaches ?
- Why is Adam preferred over SGD ?



Conceptual Questions – PCA

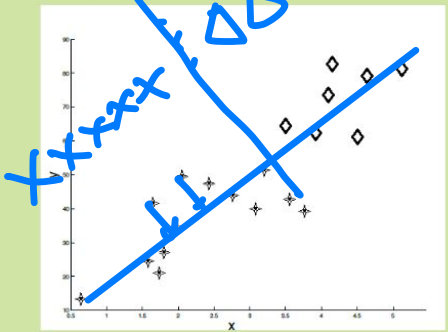
4 Principal Component Analysis [16 pts.]

(a) In the following plots, a train set of data points X belonging to two classes on $\mathbb{R}^2$ are given, where the original features are the coordinates (x, y). For each, answer the following questions:

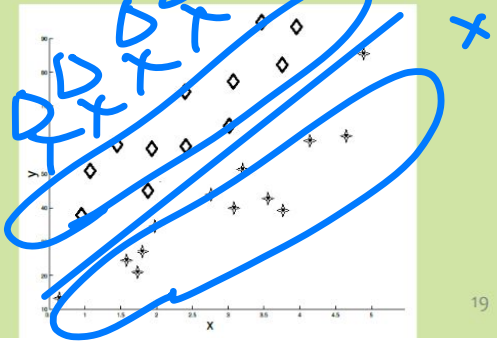
(i) [3 pt.] Draw all the principal components.

(ii) [6 pts.] Can we correctly classify this dataset by using a threshold function after projecting onto one of the principal components? If so, which principal component should we project onto? If not, explain in 1–2 sentences why it is not possible.

Dataset 1:



Dataset 2:



From 10-601 review (Matt Gormley, 2016)

Conceptual Questions – Reinforcement Learning

7.1 Reinforcement Learning

3. (1 point) **Please select one statement that is true for reinforcement learning and supervised learning.**

   ○ Reinforcement learning is a kind of supervised learning problem because you can treat the reward and next state as the label and each state, action pair as the training data.

   ○ Reinforcement learning differs from supervised learning because it has a temporal structure in the learning process, whereas, in supervised learning, the prediction of a data point does not affect the data you would see in the future.

4. (1 point) **True or False:** Value iteration is better at balancing exploration and exploitation compared with policy iteration.

   ○ True

   ○ False

From 10-601 review (Matt Gormley, 2021)

Conceptual Questions – EM

Answer TRUE or FALSE.

- The EM algorithm optimizes a lower bound on its objective function, which is the marginal likelihood $\prod_i P(x_i)$ of the observed data points $x_1, x_2, \ldots x_N$.

  TRUE. See slides 42 in lecture 10.

- The EM algorithm is guaranteed to never decrease the value of its objective function on any iteration.

  TRUE: See slides 45-50 in lecture 10.

- The objective function optimized by the EM algorithm can also be optimized by a gradient descent algorithm which will find the global optimal solution, whereas EM finds its solution more quickly but may return only a locally optimal solution.

  False. The only false part is that the gradient descent algorithm will find the global optimal solution. Gradient descent can also get stuck in a local optima.

Conceptual Question: Classification Models

Which of the following classifiers will perform best over the given dataset and why ?

- Linear SVM
- MLP
- Naive Bayes
- Perceptron
- Logistic Regression

