PCSE 595
Special Topics in
Machine Learning and Neural Networks

Dr. Sam Henry

samuel.henry@cnu.edu

Luter 325

Exam 1

Exam 1 is on Mon Feb 26th

- Math Background
  - Normal distributions, variance, covariance
  - Basic matrix operations
- Machine learning process
- Hyper parameter tuning
- Feature Engineering
- Data Normalization

- Overfitting an underfitting
  - Bias and Variance
  - Cross-validation
  - Regularization
    - L1 and L2
- Loss Functions
  - Mean Squared Error
  - Binary Cross Entropy
- Gradient Descent
  - “vanilla”, stochastic, mini-batch

- KNN Classifier
  - Distance Measures
- Nearest Centroid Classifier
- Decision Trees
  - Information Theory
- Perceptron
- Linear Regression
- Polynomial Regression
- Logistic Regression
- Artificial Neural Networks
  - Structure
  - Back propagation

Focus is on theory and reasons WHY we do things. Some focus on how.

Hyperparameters

- KNN Classifier
  - K, Distance Measures
- Nearest Centroid Classifier
  - Distance Measures
- Decision Trees
  - Splitting Criteria, Max Depth
- Perceptron
- Linear, Polynomial, Logistic Regression
  - Maybe: Learning rate, momentum
  - Regularization method
  - Function: linear, logistic, polynomial degree
  - Loss Function: Binary Cross Entropy, Sum of Squared Error (or Mean Squared Error)
- Artificial Neural Networks
  - Number of hidden neurons
  - Early Stopping and/or regularization method

Advice

- Study!
  - Review the course slides
  - Make sure you can answer the questions in these slides
  - Make sure you understand the reasoning behind the answers
- Come to me with questions

- Don’t overthink it!
  - answers can be fully explained in 1-3 sentences (more often 1 sentence)

Multi-variate Gaussian Distributions

Question

What do the diagonal elements of a covariance matrix tell us about the data?

Question

What do the diagonal elements of a covariance matrix tell us about the data?

The variance of a single dimension (the amount of variation in that dimension)

Question

What do the off-diagonal elements of a covariance matrix tell us about the data?

Question

What do the off-diagonal elements of a covariance matrix tell us about the data?

The covariance between those two features (the amount they vary relative to each other)

Question

Is a covariance matrix always symmetric?

Question

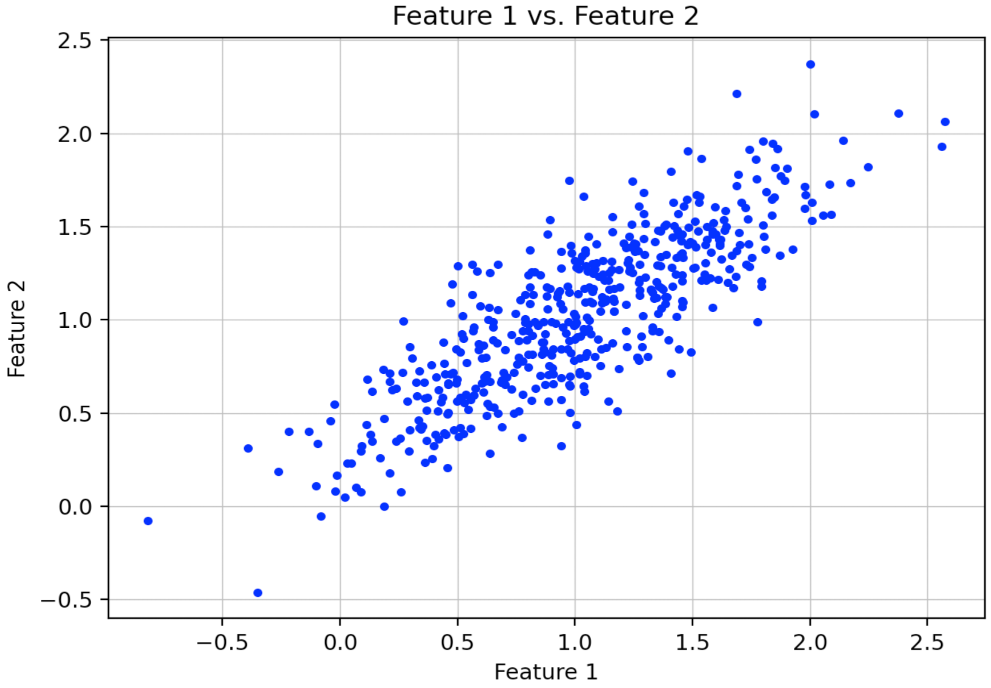
Is a covariance matrix always symmetric?

- Yes

Question

Given the following plot, which of the following is true about the features.

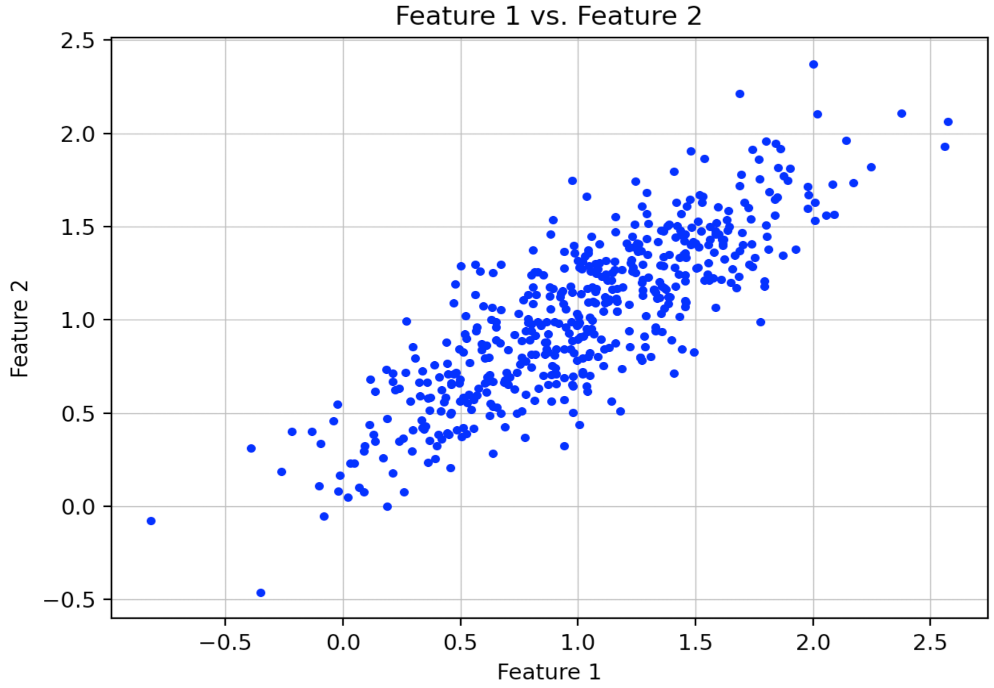
- A positive covariance
- A negative covariance
- Low/No covariance



Question

Given the following plot, which of the following is true about the features.

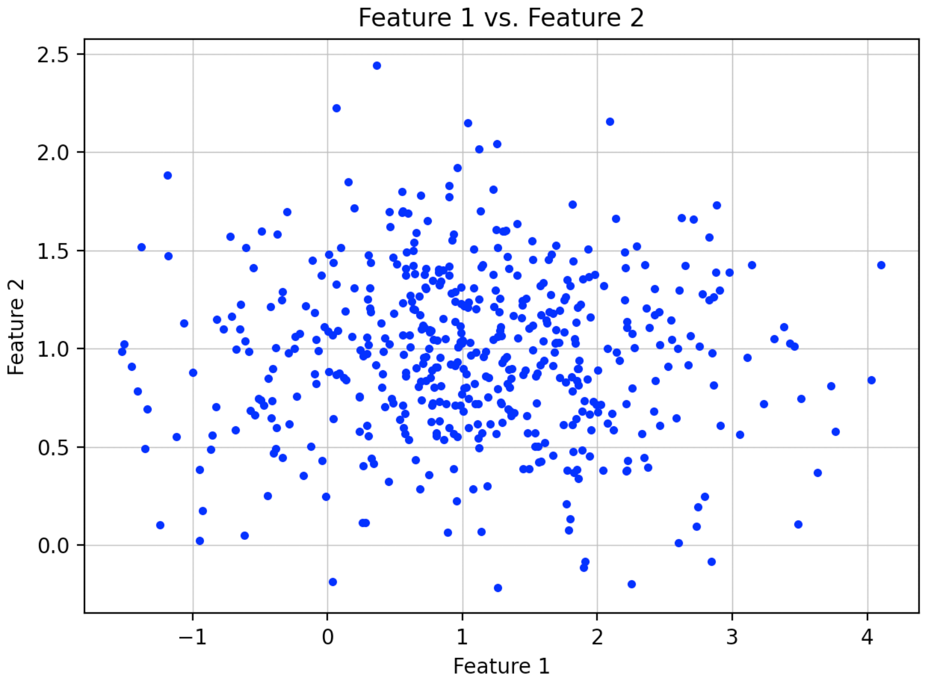
- A positive covariance
- A negative covariance
- Low/No covariance



Question

Given the following plot, which of the following is true about the features.

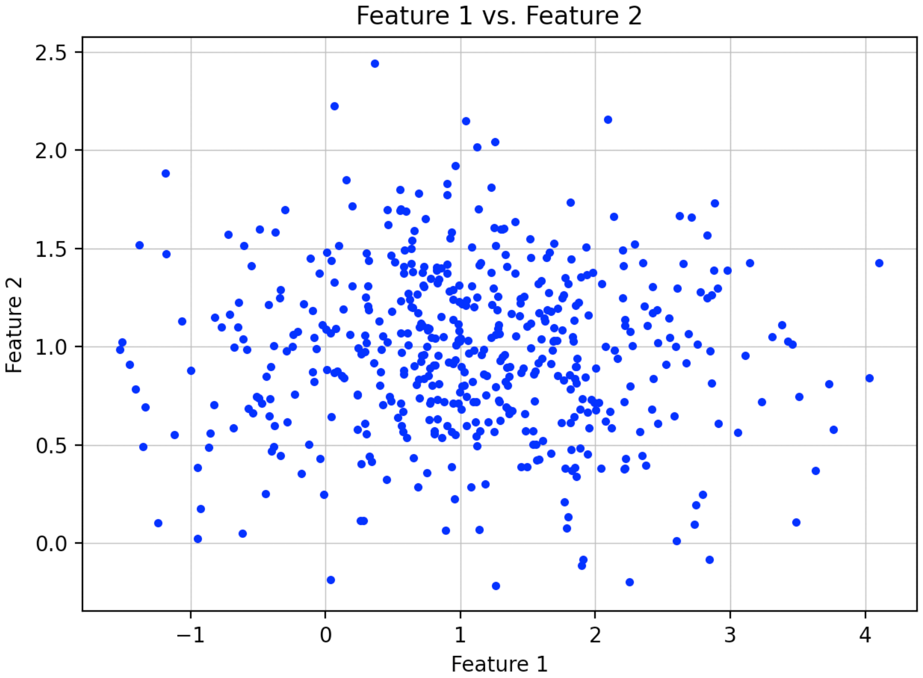
- A positive covariance
- A negative covariance
- Low/No covariance



Question

Given the following plot, which of the following is true about the features.

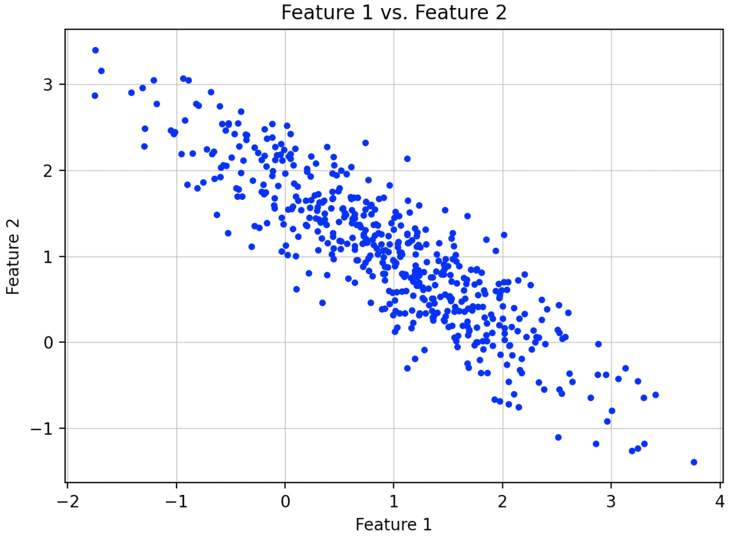
- A positive covariance
- A negative covariance
- Low/No covariance



Question

Given the following plot, which of the following is true about the features.

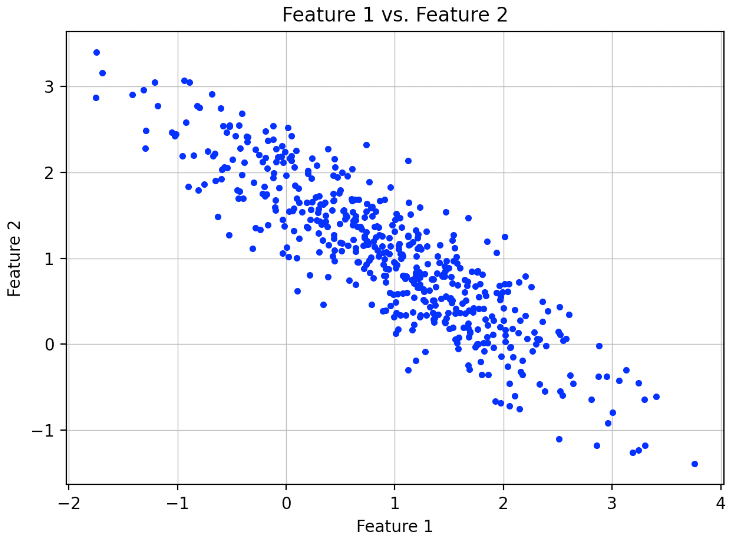
- A positive covariance
- A negative covariance
- Low/No covariance



Question

Given the following plot, which of the following is true about the features.

- A positive covariance
- A negative covariance
- Low/No covariance



Question

If feature 1 and feature 2 have a high positive covariance, what does that mean?

Question

If feature 1 and feature 2 have a high positive covariance, what does that mean?

If feature 1 increases, feature 2 increases

Question

If feature 1 and feature 2 have a high negative covariance, what does that mean?

Question

If feature 1 and feature 2 have a high negative covariance, what does that mean?

As feature 1 increases, feature 2 decreases

Question

If feature 1 and feature 2 have a covariance of zero, what does that mean?

Question

If feature 1 and feature 2 have a covariance of zero, what does that mean?

They are linearly independent, there is no relation between them

Overfitting, Underfitting, and Cross Validation

Question

What is the difference between a test set and a validation set?

Question

What is the difference between a test set and a validation set?

Test set is held out during model development. It is used to evaluate the system after all hyper-parameters have been tuned. Validation set is held out during model training, but used to evaluate generalizability during hyper-parameter tuning and model selection.

Question

How does the bias and variance trade-off relate to model complexity?

Question

How does the bias and variance trade-off relate to model complexity?

- Simpler models = more bias, less variance
- complex models = more variance, less bias

Question

How would you decrease the potential for overfitting in a k-nearest neighbors algorithm?

Question

How would you decrease the potential for overfitting in a k-nearest neighbors algorithm?

- Increase k

Question

How would you decrease the potential for overfitting in a polynomial regression?

Question

How would you decrease the potential for overfitting in a polynomial regression?

- Choose a smaller degree
- Add regularization

Linear, Polynomial, and Logistic Regression

Question

How does an objective function relate to the weights of our model?

Question

How does an objective function relate to the weights of our model?

- The objective function determines the loss for a particular choice of weights
- We can plot it, and it is a plot of model weights versus the loss of the corresponding model

Question

Why do we want to find where the gradient is zero in our objective functions?

Question

Why do we want to find where the gradient is zero in our objective functions?

- That is the minimum value of the objective function, which tells us the optimum values of model weights

Question

How do you find an analytical solution for an objective function that is convex? What is the intuition behind the solution?

Question

How do you find an analytical solution for an objective function that is convex? What is the intuition behind the solution?

- Take the derivative, set the equation equal to 0, and solve for theta. This is equivalent to finding the values of theta that minimize the objective.
- A convex function has a single global minimum which is where the derivative is zero. (Draw parabola)

Question

Why do we want our objective function to be convex?

Question

Why do we want our objective function to be convex?

- So that there is a global minimum

Question

What is binary cross entropy and when is it used?

Question

What is binary cross entropy and when is it used?

- Binary cross entropy is a loss function. It is primarily used as a loss function for binary classification

Question

What is mean squared error and when is it used?

Question

What is mean squared error and when is it used?

- Mean squared error is a loss function. It is primarily used as a loss function for regression

Gradient Descent

Question

- During training, what is an epoch?

Question

- During training, what is an epoch?
  - An iteration over the entire dataset

Question

- What are you “descending” in gradient descent?

Question

- What are you “descending” in gradient descent?
    - You descend the objective/loss/error function (to find the minimum).

Question

- How do you know what direction to descend?

Question

- How do you know what direction to descend?
  - You move in the direction of the negative gradient

Question

- What does the end result of gradient descent tell you?

Question

- What does the end result of gradient descent tell you?
  - The model weights that (ideally) minimize the objective function (minimize loss)

Question

What is the difference between standard (“vanilla”) gradient descent and mini-batch gradient descent?

Question

What is the difference between standard (“vanilla”) gradient descent and mini-batch gradient descent?

Standard gradient descent calculates loss over the entire dataset before updating. Mini-batch gradient descent calculates loss over a subset of the data (a batch) before updating

Question

Why might we use mini-batch gradient descent instead of standard (“vanilla”) gradient descent?

Question

Why might we use mini-batch gradient descent instead of standard ("vanilla") gradient descent?

- Mini-batch gradient descent is useful when you have a large dataset, because iterating over the whole dataset to estimate the error takes a long time, and you can get a 'good enough' estimate with a batch-size. Therefore allowing more steps to be made in a faster amount of time.

Question

How is stochastic gradient descent different from standard (“vanilla”) gradient descent?

What are the advantages and disadvantages?

Question

How is stochastic gradient descent different from standard (“vanilla”) gradient descent? What are the advantages and disadvantages?

- Stochastic gradient descent estimates the error with a single sample, and takes a step after each sample. This is good because you can take a lot of steps during training; this is particularly useful for large datasets. However, it is bad because your estimate of error (and therefore the derivative, which is the direction you move in) is inaccurate. Therefore, you may move in the wrong direction.

Question

Why do we use a learning rate during gradient descent?

Question

Why do we use a learning rate during gradient descent?

The learning rate scales the step size (weight update).

When the gradient is very large, it can prevent us from overstepping the minimum.

When the gradient is very small, it can help us arrive at a good solution faster

Regression

Question

Why do we add a bias when performing regression?

Question

Why do we add a bias when performing regression?

- It shifts our hyperplane off of 0.
- It controls the height of our hyperplane.
- …this is important because we don’t assume our data is centered at 0

Question

Why don’t we assume a Bernoulli distribution in regression problems?

Question

Why don’t we assume a Bernoulli distribution in regression tasks?

Bernoulli distributions assume values are one of two values, but in regression tasks values are real-valued.

Question

What is the relationship between maximum likelihood estimation and binary cross entropy?

Question

What is the relationship between maximum likelihood estimation and binary cross entropy?

- We derived binary cross entropy using the maximum likelihood estimation of Bernoulli distributed data

Question

Why is linear regression an inappropriate method for solving classification problems?

Question

Why is linear regression an inappropriate method for solving classification problems?

- Linear regression assumes there is a linear relationship between the dependent and independent variables, which is not the case for classification problems

Question

Given that b indicates the bias, x_{01} indicates the value of sample 0 dimension 1, x_{12} indicates the value of sample 1 dimension 2, and so on

Write the matrix that results from applying a 3rd degree polynomial basis function to the following data matrix. (I expect values to be written as an expression/equation)

$$\begin{bmatrix} b & x_{01} & x_{02} \\ b & x_{11} & x_{12} \end{bmatrix}$$

Question

Given that b indicates the bias, x_{00} indicates the value of sample 0 dimension 1, x_{12} indicates the value of sample 1 dimension 2, and so on

Write the matrix that results from applying a 3rd degree polynomial basis function to the following data matrix. (I expect values to be written as an expression/equation)

$$\begin{bmatrix} b & x_{01} & x_{02} \\ b & x_{11} & x_{12} \end{bmatrix}$$

$$\begin{bmatrix} b & x_{01} & x_{02} & x_{01}^2 & x_{02}^2 & x_{01}^3 & x_{02}^3 \\ b & x_{11} & x_{12} & x_{11}^2 & x_{12}^2 & x_{11}^3 & x_{12}^3 \end{bmatrix}$$

Question

Since high degree polynomials can fit more data distributions, why wouldn’t you always use a high degree polynomial to fit data?

- They are prone to overfitting
- There may be numerical instability issues with very high degrees

Regularization

Question

- What is the purpose of regularization?

Question

- What is the purpose of regularization?

  To avoid overfitting

Question

- Describe at a high level how regularization works

Question

- Describe at a high level how regularization works
  - Regularization attempts to minimize model weights. The purpose is to avoid overfitting (under the assumption that higher the weight values, the more complex the model which are therefore more prone to overfitting).

Question

- Why would regularization try to minimize model weights?

Question

- Why would regularization try to minimize model weights?

Many regularizers assume that higher the weight values, the more complex the model which are therefore more prone to overfitting

Question

What is the difference between L1 and L2 regularization?

Question

What is the difference between L1 and L2 regularization?

- L1 regularization is the sum of absolute value of weights
- L2 regularization is the sum of squared weights

Information Theory
and Decision Trees

Question

How do decision trees handle numeric data? You may give an example to explain your answer.

Question

How do decision trees handle numeric data? You may give an example to explain your answer.

- By finding split point between classes, and thereby converting the data into binary features.

True or False

A major benefit of decision trees is that they are highly interpretable

A major benefit of decision trees is that they can handle categorical data

Information depends on context and surprise

A KNN classifier makes few assumptions about the distribution of the data

True or False

A major benefit of decision trees is that they are highly interpretable

True

A major benefit of decision trees is that they can handle categorical data

True

Information depends on context and surprise

True

A KNN classifier makes few assumptions about the distribution of the data

True

Question

How is information gain used in decision trees?

Question

How is information gain used in decision trees?

- Information grain is used to determine the optimum features to split on in a decision tree

Question

What is the purpose of a max depth hyperparameter in decision trees?

Question

What is the purpose of a max depth hyperparameter in decision trees?

- To prevent overfitting

Question

Given the following equations for entropy

Calculate the entropy of the following dataset.

$$\widehat{H}(X) = -\sum_{i=1}^{c} \widehat{p_i} * log_2(\widehat{p_i})$$

Features	Class
x_1= <t,d>	y_1= +
x_2= <s,d>	y_2= +
x_3= <t,b>	y_3= -
x_4= <t,r>	y_4= -
x_5= <s,b>	y_5= +

Question

Given the following equations for entropy

Calculate the entropy of the following dataset

(Just expand the equation an plug in values. You don’t need to compute logarithms)

Features	Class
x_1= <t,d>	y_1= +
x_2= <s,d>	y_2= +
x_3= <t,b>	y_3= -
x_4= <t,r>	y_4= -
x_5= <s,b>	y_5= +

$$\widehat{H}(X) = -\sum_{i=1}^{c} \widehat{p_i} * log_2(\widehat{p_i})$$

$$H(X) = -\frac{3}{5} * log_2\left(\frac{3}{5}\right) - \frac{2}{5} * log_2\left(\frac{2}{5}\right) = 0.971 \text{ bits}$$

Question

If a dataset has an entropy of 1, what does that tell you?

Conversely, if a dataset has an entropy of 0, what does that tell you?

Question

If a dataset has an entropy of 1, what does that tell you?

- There is an even class distribution in the dataset, meaning we will need exactly one bit (heads or tails) to communicate an outcome

Conversely, if a dataset has an entropy of 0, what does that tell you?

- Everything in the dataset is the same class, meaning we will need zero bits to communicate an outcome. Since the outcome is always the same, there is no surprise, and therefore no information)

Artificial Neural Networks

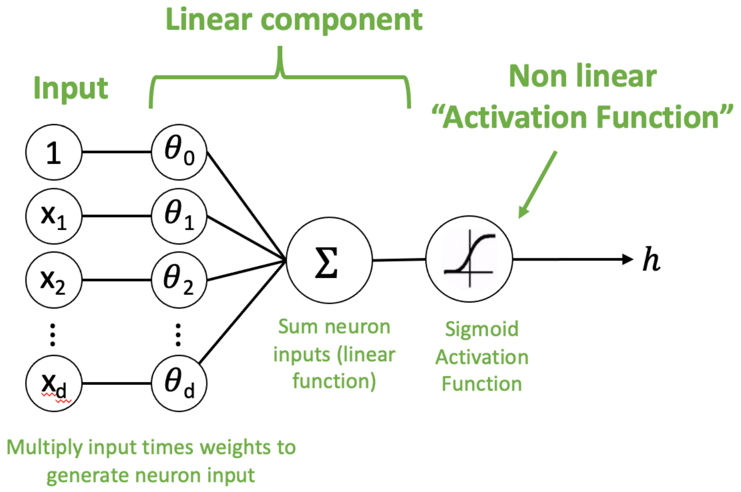
Question

What is an artificial neuron? Describe it pictorially and label all components.

What is the function it computes?

Question

What is an artificial neuron? Describe it pictorially and label all components.


Linear component
Input
Non linear
"Activation Function"
1
x1
x2
xd
θ0
θ1
θ2
θd
Σ
h
Sum neuron inputs (linear function)
Sigmoid Activation Function
Multiply input times weights to generate neuron input

What is the function it computes?

$$h = \frac{1}{1 + e^{-\sum_{i=1}^{d} \theta_{i*} x_i}}$$

Question

- What does it mean that an ANN is a universal approximator?

Question

- What does it mean that an ANN is a universal approximator?

That it can approximate any function arbitrarily well

Question

- How does the number of neurons in the hidden layer relate to bias and variance?

Question

- How does the number of neurons in the hidden layer relate to bias and variance?

More neurons = more inflection points = more potential to overfit = higher variance, lower bias
Less neurons = less inflection points = less potential to overfit = lower variance, higher bias

Question

What is backpropagation?

Question

What is backpropagation?

Short Answer (which is OK):
The algorithm typically used to update weights in a neural network.

Longer Explanation:
It is an algorithm to efficiently compute the derivative of each weight in a neural network. It is derived primarily using chain rule. The derivative of each weight is used to update it during training, which is typically gradient descent. In which case, each weight is updated by moving in the direction of the negative derivate scaled by the learning rate.

Question

- Can I use regularization with a neural network?

Question

- Can I use regularization with a neural network?

Yes, this is often called “weight decay” for neural networks

Question

- Can I use a neural network for regression? If so, how?

Question

- Can I use a neural network for regression? If so, how?

Yes, use a linear activation function on the output layer

Question

- Given that the update formula for a weight is:

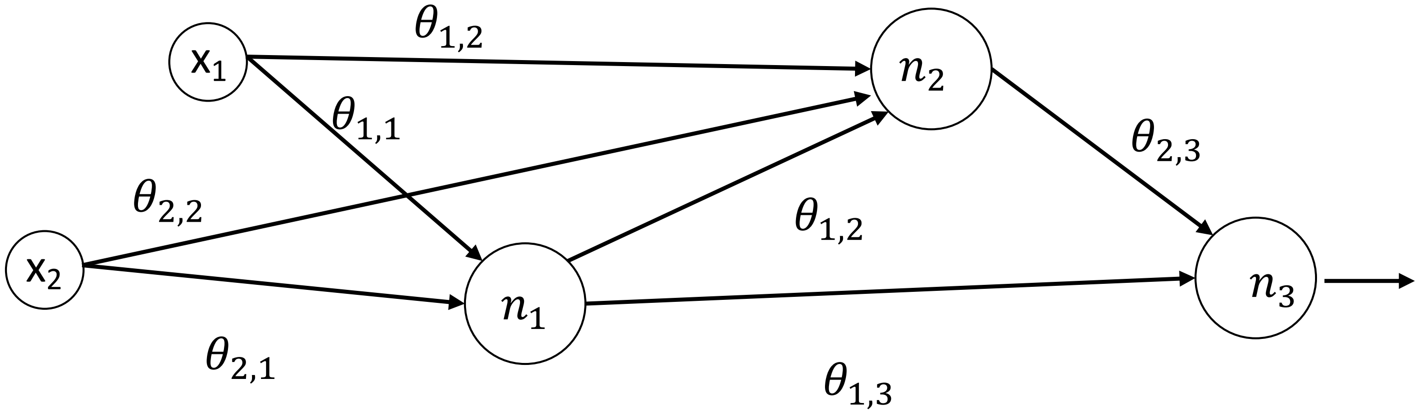
$$gradient_j = z_i * h_j(1 - h_j) * \left(\sum_{m=1}^{m} \theta_{k,m} * e_m \right)$$

Where z_i is in the input of a neuron, h_j is the output of a neuron, and e_m is the error of neuron m connected via path with weight $\theta_{k,m}$

And:

$$error\ of\ neuron\ 3, e_3 = \hat{y} - y$$

Given the following neural network with neurons 1, 2, and 3, show the equation for the weight update for $\theta_{1,1}$



Question

$$gradient_{1,1} = x_1 * h_1(1 - h_1) * \left(\theta_{1,2} * \left(\theta_{2,3} * (\hat{y} - y)\right) + \theta_{1,3} * (\hat{y} - y)\right)$$

- Given that the update formula for a weight is:

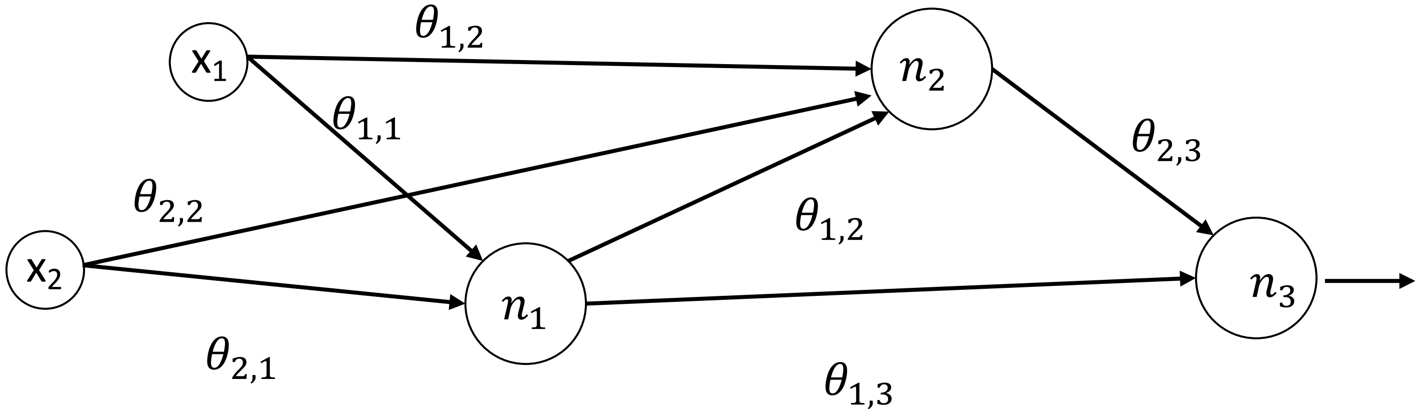
$$gradient_j = z_i * h_j(1 - h_j) * \left(\sum_{m=1}^{m} \theta_{k,m} * e_m\right)$$

Where z_i is in the input of a neuron, h_j is the output of a neuron, and e_m is the error of neuron m connected via path with weight $\theta_{k,m}$

And:

$$error\ of\ neuron\ 3, e_3 = \hat{y} - y$$

Given the following neural network with neurons 1, 2, and 3, show the equation for the weight update for $\theta_{1,1}$



Question

- My neural network has a logistic activation function in its final layer. The network is trained to predict if emails are Spam or Ham (not spam). I encode my classes as 1 = Spam, 0 = Ham
- Given a sample, the network outputs a 0.6.
- What is the probability that the sample is Ham?

Question

- My neural network has a logistic activation function in its final layer. The network is trained to predict if emails are Spam or Ham (not spam). I encode my classes as 1 = Spam, 2 = Ham
- Given a sample, the network outputs a 0.6.
- What is the probability that the sample is Ham?

1 - 0.6 = 0.4

Perceptron

Question

- Is a perceptron a linear classifier?

Question

- Is a perceptron a linear classifier?

Yes

Question

- How does a perceptron deal with non-linearly separable data?

Question

- How does a perceptron deal with non-linearly separable data?

It doesn’t. A perceptron will not converge to a solution for non-linearly separable data.
Furthermore, there is no guarantee that a perceptron’s decision boundary gets better over time, so adding a max_epoch’s parameter can stop it, but the final answer may not be any good.

KNN

Question

- What is “K” in a K-Nearest Neighbors Classifier, and what effect does changing it have? Relate your answer to bias and variance

Question

- What is “K” in a K-Nearest Neighbors Classifier, and what effect does changing it have? Relate your answer to bias and variance

The K closest points to an unlabeled point “vote” on the class of a new sample.

Large values of K decrease variance and increase bias

Small values of K increase variance and decrease bias