This page has questions **without answers**.
With Answers Here

Fundamentals of Learning

Lecture 2-3: K-NN, Classification, Regression, and Data

1. We want to estimate our model's error when trained on N training samples, using a test set with M samples. True or false:

a) With different sets of M samples, we would probably get the same error measurement.

b) If we increase M, we should get a more accurate estimate (i.e. one that has lower variance with different test sets of size M)?

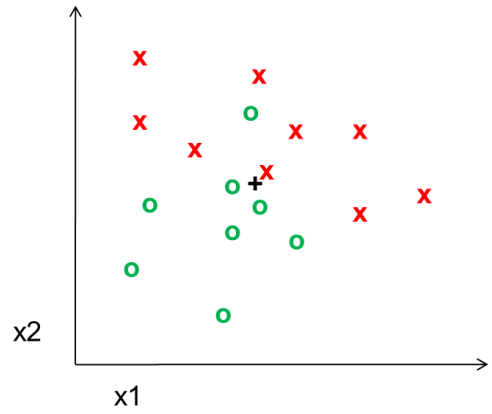
c) If we increase N but do not change the test set, we'd expect the test error to be unchanged.

d) The expected error does not depend on M, but it does depend on N.

2. Which of these assumptions are implied by the use of a Euclidean or L2 distance measure for a KNN classifier? (can select any number)

a. Each feature dimension is equally important
b. The feature dimensions have comparable scales
c. Each feature dimension has roughly the same mean

3. Classify ('o' or 'x') the '+' with 1-NN and 3-NN



1-NN:
3-NN:

4. Which of these are true of nearest neighbor? (choose all that apply)

a. Fast inference

b. Fast training
c. Can be applied if only one sample per class is available
d. Is not commonly used in practice
e. Is most powerful when used in combination with feature learning

Lecture 4: Clustering and Retrieval

1. Indicate whether the following is true or false. If something is false, indicate what is wrong with the statement justifying your response appropriately.

a) The basic idea of K-means is to assign each point to the nearest of the established K centers.

b) The problem with a very structured distribution of points is that it can make the K-means algorithm not converge to a final clustering of all the data points.

c) High-dimensional data points, even if we have a small number of them, compared to low-dimensional data points, make K-means (with the same number of clusters) iterate more times before it achieves a sufficiently good clustering.

d) High-dimensional data points increase the computational cost spent when running K-means, and thus people in practice just stop it before even achieving a “stable” clustering.

e) K-means is a deterministic algorithm, but it is sensitive to the initialization of the centers for each cluster.

f) If we don’t know much about the data, in practice, people determine the number of clusters for K-means based on memory or computational requirements.

g) The problem with clustering methods such as K-means and hierarchical K-means is that they can be sensitive to the local connectivity of the data.

i) If we know that some attributes are more important than others, we can still use K-means and expect a good clustering.

j) One big advantage of hierarchical K-means is that it is computationally more efficient.

k) Agglomerative clustering can be sensitive to the local connectivity of the data, as long as we have a good choice of linkage function.

l) The idea of LSH is used in those situations where we are not really concerned about retrieving the “closest point” to a query, but we are happy to have “a sufficiently close enough point” to the query.

2. If you have vectors with continuous values that you want to somehow group into different categories, how do clustering methods help?

3. If you have a group of pictures, which attributes could be used in order to produce a good clustering?

4. Imagine you have a group of unlabelled or "non-annotated" data and you used two clustering algorithms (with the same final number of clusters) to separate the data into groups. How would you compare the results of both clustering methods?

5. For each of the following statements, select the distance measure that best corresponds, from L2, L1, Mahalanobis

a. Each dimension has the same scale, distance can be dominated by large differences in one dimension
b. Each dimension has the same scale, sensitive to the sum of absolute differences

Lecture 5: Dimensionality Reduction: PCA and Low-D Embeddings

1. Indicate whether the following is true or false. If something is false, indicate what is wrong with the statement justifying your response appropriately.

a) When doing dimensionality reduction, the idea is that the points in the lower dimension should preserve "some" relationship they had in their original dimension.

b) PCA is based on computing eigenvectors of the empirical covariance matrix, but a problem (which seldomly happens in practice) is that sometimes such eigenvectors can have imaginary entries and make PCA useless.

c) When using PCA, the eigenvectors of the empirical covariance matrix are able to capture discriminative features from the data and preserve them in the lower dimensional representation.

d) Depending on the application, the components computed using PCA may have a particular qualitative significance, as in the case of eigenfaces.

e) In every application of PCA, the largest components (the ones associated to the largest eigenvalues) are always the most important to preserve.

f) Non-linear embedding methods and manifold learning focuses on representing the relationships between data points, even though reconstruction may not be possible (unlike PCA).

g) MultiDimensional scaling (MDS) aims to preserve pairwise distances between points in the lower dimension, where the distance can be arbitrarily defined by the user according to the application.

h) A dataset typically obtained for psychology or marketing applications may have its data points related by ratings, opinions and even feelings (e.g., think of pairwise ranking of multiple objects based on the preferences from an individual). MDS is guaranteed to be used in this case since we can always define an appropriate distance metric.

i) Dimensionality reduction based on ISOMAP defines a unique graph over which we calculate the distance between two points based on the shortest path between them.

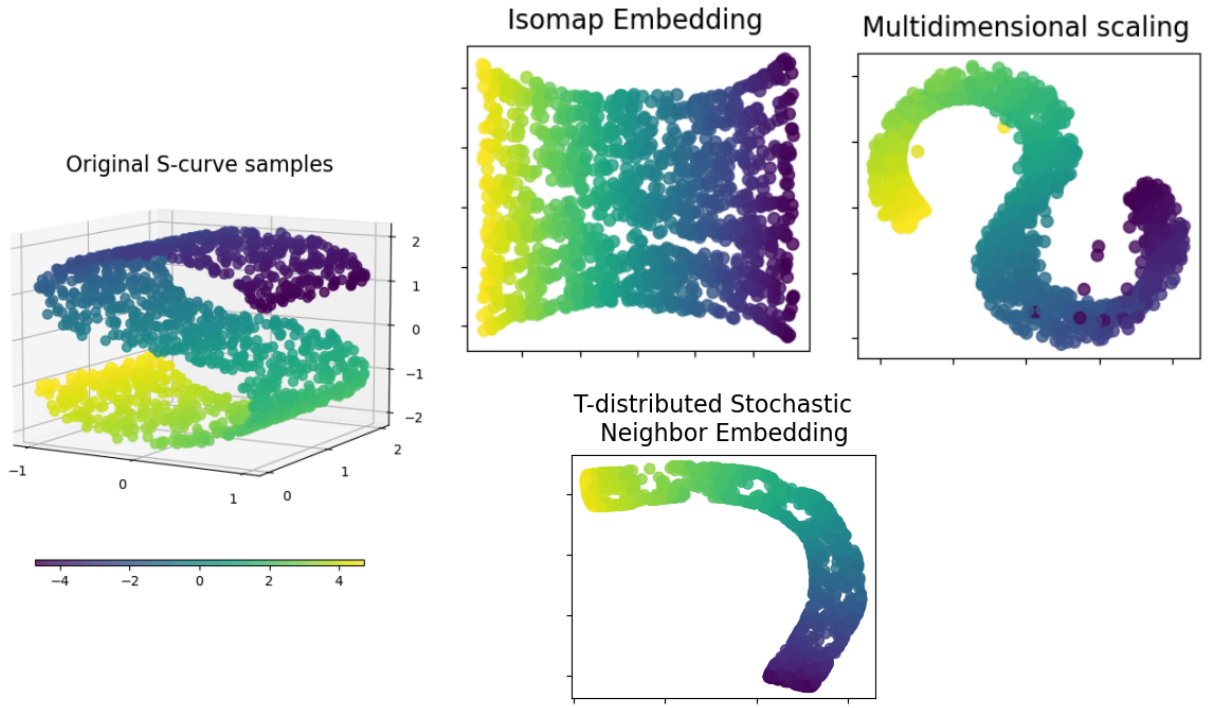
j) t-SNE computes pairwise probability distributions for the data in its original dimension (or old coordinates) and for the lower dimension (or new coordinates). The objective is then to find the new coordinates that minimize the KL divergence between those two probability distributions.

k) UMAP is computationally less expensive than other non-linear scaling or manifold learning methods, which makes it widely used in practice.

2. How would you choose the number of components when using PCA?

3. When facing a high-dimensional dataset, a researcher who wants to use MDS first uses PCA as a preliminary step to reduce some of the dimensions. Provide two reasons that could explain why this is the case.

4. In one of the slides in class we observed the following example of dimensionality reduction by three methods.



https://scikit-learn.org/stable/auto_examples/manifold/plot_compare_methods.html

a) Why does MDS have an S-shape?
b) How would you explain the fact that t-SNE does not have an S-shape?

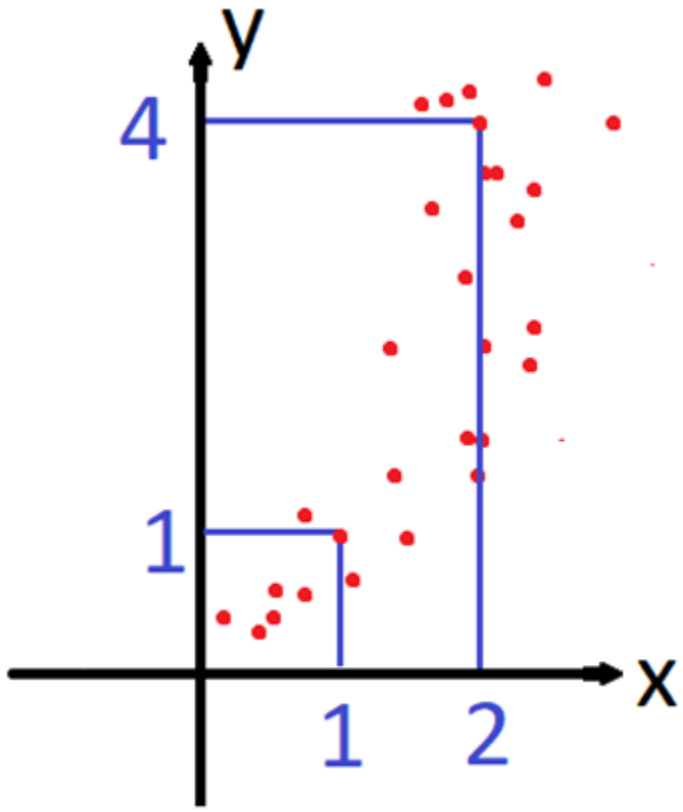
Lecture 6 and 7: Linear Regression, Regularization, Logistic Regression, SVM

1. How do L1 and L2 regularization complement each other in linear models (logistic regression or linear regression)?

2. For each case, does it make more sense to use logistic regression (LG) or linear regression (LR)? Why?

a) Predict the likelihood that a company's stock price is overvalued based on recent historical accounting features
b) Predict the future earnings of a company based on past historical performance
c) Predict know whether the current inflation increase will drastically, mildly, or lightly decrease the average price of the stock market

3. **Someone shows you the following plot**



and claims that it is impossible to use linear regression *by any means* since the red dots (data) do not follow a linear trend (and assume for the sake of the question that indeed they do not follow any linear trend). Is this person right or not? If not, what could he do to use linear regression?

4. Indicate whether each statement is TRUE or FALSE and explain why:

a) One method of hyperparameter tuning is to use a special transformation of variables which could let us use the training data to optimize for the hyperparameter.

b) Cross-validation splits the training data so that we can measure the performance of the hyperparameter over different splits using one portion of each as a validation set

c) When using linear regression or logistic regression with sufficient data there is no need to have regularization.

5 Linear regression is based on the minimization of squared differences, which makes it very vulnerable to outliers. How would you (informally) define an outlier? Can you graphically provide a qualitative explanation of how outliers can be detrimental to linear regression?

1. Indicate whether the following is true or false. If something is false, indicate what is wrong with the statement justifying your response appropriately.

a) SVMs are more explainable than artificial neural networks.

b) The dual representation of SVMs provides further insights into linear classifiers, showing that optimal parameters correspond to a non-linear combination of training examples.

c) When training a SVM we minimize the margin in order to improve generalization.

d) Unlike SVM, linear logistic regression always adds a non-zero penalty over all training data points. True

e) Hinge loss increases quadratically depending on how far is the misclassified data point from the decision boundary, and gives a zero loss for a correctly classified one.

f) Representation theorem states that, in the context of linear classifiers (with L2 penalty), it is impossible to represent the optimal parameter by a linear function of the training data.

g) An advantage of SVMs is that we can use the so-called kernels during the learning process to implement feature mapping without having to perform an explicit mapping over each datapoint.

h) Using a soft margin allows the training of SVMs which tolerate the misclassification of some points with the purpose of maximizing the margin. True

i) A general disadvantage in the soft margin is the hindering of generalization.

j) A radial basis function SVM is good at classifying data points where points of one class form some sort of ellipsoid cluster, while the ones from the other class are outside around such a cluster.

k) Removing a support vector from the training data (after learning an SVM on it) can affect the size of the margin but will never affect the decision boundary.

2. Why is it that SVMs don't depend on the whole training set? What are two advantages of this feature?

3. Why are we more likely to find the decision boundary of linear logistic regression farther from dense clusters of training data points compared to an SVM?

Lecture 8: Probability and Naïve Bayes

Note: some of this may be from the Probability Tutorial

1. What assumption does the Naive Bayes model make if there are two features x1 and x2?

(a) $P(y|x1, x2) = P(y|x1)P(y|x2)$
(b) $P(x1, x2|y) = P(x1|y)P(x2|y)$
(c) Neither of these are true
(d) These are equivalent and both true

2. Which of these are true?

(a) If you want to solve for the likelihood parameters in Naive Bayes, you need to consider many features at once
(b) A continuous feature can be modeled as a Gaussian, or converted into discrete values and modeled as a multinomial
(c) Naive Bayes is guaranteed to underperform nearest neighbor, since Naive Bayes makes stronger assumptions
(d) Compared to most ML algorithms, Naive Bayes is relatively fast to train and fast to predict

3. True or False: P(a|b) = P(b|a)

4. True or False: P(a,b) = P(b,a)

5. True or False: a and b are independent

P(a=0,b=0) = 0.12
P(a=0,b=1) = 0.08
P(a=1,b=0) = 0.48
P(a=1,b=1) = 0.32

6. True or False: If x1 is independent of x2, then x1 and x2 are conditionally independent, given y

7. According to Bayes rule, P(y|x) =

(a) P(x|y)P(y)/P(x)
(b) P(x|y)P(x)/P(y)
(c) P(x,y)/P(x)/P(y)

8. Which of these are equivalent to argmax_x f(x)? (choose all that apply)

(a) argmax_x [f(x) + y]
(b) argmax_x [1/f(x)]
(c) argmax_x [log f(x)]
(d) argmax_x [exp f(x)]

9. True or False: If a prior is not used when estimating likelihood parameters, it is possible that P(x|y)P(y)=0 for all y.

Lecture 9: EM and Latent Variables

1. Indicate whether the following is true or false. If something is false, indicate what is wrong with the statement justifying your response appropriately.

a) Computing local sensitivity hashing (LSH) through random projection has the problem of producing sparse hash keys when used with high-dimensional data.

b) Generating longer hash keys using LSH by random projection can have better accuracy of retrieval (or recall) but a downside is that it can take more time to query.

c) Latent variables may refer to variables that affect the data but which cannot be directly observed from the data.

d) The EM algorithm provides a recipe for modeling latent variables.

e) In the bad annotator problem (where annotators rank pictures, and among which some bad annotators just give rankings sloppily), bad annotators may be modeled as providing ratings from a uniform distribution because their rankings are not informational.

f) Given some estimate of the latent variables and model parameters, the objective of the E-step in the EM algorithm is to estimate the likelihood of the observed data.

g) The M-step in the EM algorithm is basically solving for an MLE problem where the data and latent variables are “fixed” (coming from the E-step) and we try to come up with the value of model parameters that most likely generated them.

h) Generally, in the M-step, the estimation of the model parameters is “weighted” by the latent variable likelihoods.

i) The EM algorithm is guaranteed to always converge to the global (or best) maximum for the parameters’ values of the problem at hand.

j) In the bad annotator problem, the observed robustness of the EM algorithm comes from the fact that we can always distinguish which annotators are the “bad ones” independently from how we model them.

k) K-means is an example of a hard EM algorithm.

l) The EM algorithm is a method for maximum likelihood estimation in the presence of missing or incomplete data.

m) The EM algorithm guarantees convergence to the global maximum of the likelihood function for any given dataset.

n)The E-step of the EM algorithm computes the maximum likelihood estimate of the parameters given the observed data. False. That is closer to the M-step.

o) True or False: In the EM algorithm, the likelihood of the observed data increases after each iteration of the algorithm.

2. Given binary data x~Bernoulli, $P(x) = p^x(1 - p)^{1-x}$, what is the MLE estimate of p given $x_1 .. x_N$?

Lecture 10: Density estimation: MoG, Hists, KDE

1. Indicate whether the following is true or false. If something is false, indicate what is wrong with the statement justifying your response appropriately.

a) A parametric model for estimating PDFs is by discretization, e.g., using histograms, because the number of discretized values (or bins) are the parameters to specify.

b) A continuous random variable – described by a PDF – has probability zero of taking any single value.

c) Intuitively, a one dimensional PDF is “smooth” when small changes around a specific value x can be approximated by the value at x.

d) Using histograms to describe a PDF works better the higher the dimension of the data.

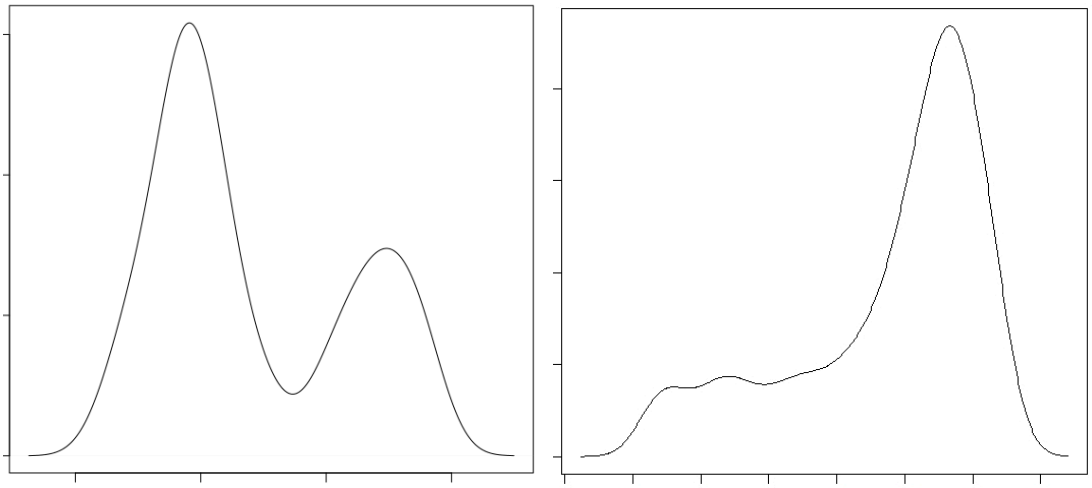
e) Mixture of Gaussians works better when the PDF that we try to approximate is smooth.

f) A beta distribution allows us to only model unimodal distributions whose skewness, width, and shape itself is parameterized by two shape parameters (alpha and beta).

g) The hyperparameters that define multiple methods to estimate PDFs (e.g., the bandwidth for the kernel density method, the number of components of the mixture of Gaussians, etc.) can be selected through cross-validation.

h) A practical assumption when estimating the PDF of a random vector is by assuming that each of its components are independent: we estimate the density of each component separately, and then define the approximated density value of the random vector as the product of the density values of each entry.

2. Assume that the following two plots describe two different PDFs.



a) If we want to use a Gaussian distribution to model such distributions, which of the two plots will be approximated better? Why?

b) Which method would you consider best to approximate the PDF on the left? Why?

3. If in theory a mixture of Gaussians is a general purpose PDF estimator because it is able to model any probability density function provided that we use enough components, why is it that in practice it may not be feasible to use this method when the PDF is not smooth and/or is more convoluted in its shape?

Lecture 11: Outliers and Robust Estimation

1. Indicate whether the following is true or false. If something is false, indicate what is wrong with the statement justifying your response appropriately.

a) Moving average allows us to eliminate or filter out any type of additive noise.

b) Outliers can take very extreme values with respect to the range of values where the signal mostly lives in. Thus, a moving average is not robust to outliers because their computed averages can be pulled far in different directions.

c) The problem with outliers is that they always represent incorrect values of the variables we are trying to characterize.

d) The median filter is a robust filter because it is not affected by extreme values.

e) The moving average has a computationally less expensive implementation than the median filter as they both grow in size.

f) The only way we can use PCA to detect outliers is by reducing the dimension of the data to two dimensions and then visually removing those points that are far from observed clusters.

g) Robust least squares deals with outliers by “saturating” the loss function it uses in its objective function.

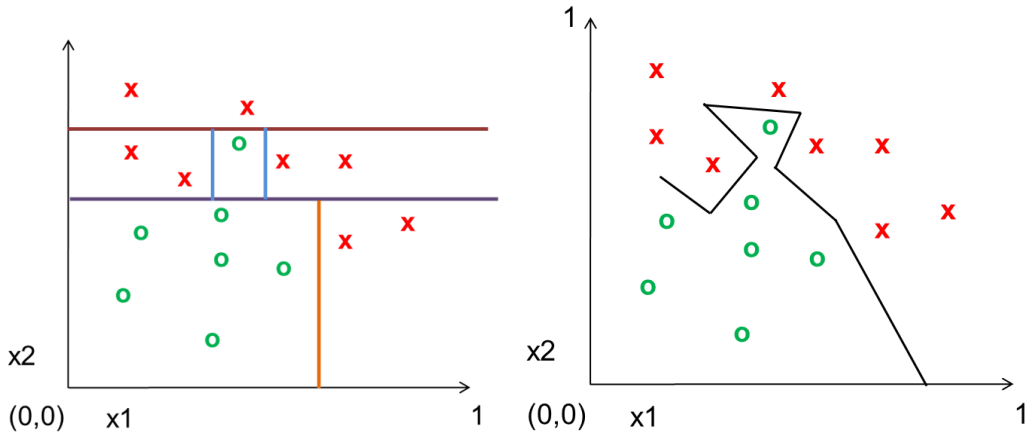
h) RANSAC is an algorithm especially designed to add robustness to the least squares method.

i) We choose the number of samples (but not the samples themselves) that defines the set over which the RANSAC algorithm will fit the model at each iteration.

2. The following image is easily found on the internet because of its wide use on the digital image processing community, it is called “Lena”:

The following two images were obtained from the thesis “On GPU-Accelerated Fast Direct Solvers and Their Applications in Image Denoising”. Both images are noisy. Which kind of filter would you use to remove the noise in each one and why?

Lecture 12: Decision Trees



1. In class we observed the following two graphs regarding the decision boundaries for decision trees (left) and 1-nearest neighbor (right). Each data point consists of values for two features x1 and x2.

a) Why are the decision boundaries "parallel" to one of the axes in this case?

b) The displayed decision boundaries of the 1-NN are lines that are not parallel to either of the axes. In general, we can also have these type of "non-parallel" lines as boundaries for decision trees, how could this happen?

c) As a follow up to the previous question: in a practical setting, what could be a possible disadvantage of using decision boundaries that are not "parallel" to the axes in decision trees?

2. Can any type of decision tree always be expressed as a binary tree? Why?

3. An important part of the algorithm studied in class for building a node in a decision tree is to choose an attribute and then split the given data for the node based on the values of the attribute.

a) Do all attributes in a decision tree have to be of the same type (discrete/categorial or continuous)? Or is it possible for them to be of different types? In addition, provide an example to support your answer.

b) How do we choose the "best" attribute depending on whether its values are discrete (i.e., categorical) or continuous?

4. Why is it a good decision to use early stopping during the training of decision trees?

5. Assume we have the following table, each row being a data point:

x1	x2	y
T	F	T
F	F	T
T	F	F
T	T	T
F	T	T

F	F	T

Compute the H(y), H(y|x1) and H(y|x2). If you need to choose whether to split this data based on x1 or x2, which feature should you choose according to information gain criterion?

Lecture 13: Ensembles and Random Forests

1. Indicate whether the following is true or false. By supervised learning we mean the learning from labeled data in either classification or regression tasks.

a) Underfitting happens when there is scarce training data which does not allow the model to generalize well.

b) Model complexity depends on the quality of the data – for example, better quality data would fit better a given specified model, thus reducing the complexity in the training of the model.

c) The idea of boosting is to combine multiple weak classifiers to obtain a strong one.

d) Bagging is used to lower the variance of predictor models over a dataset.

e) Random forests is a type of bagging.

f) AdaBoost algorithms such as Real AdaBoost gives more weights to those samples that have been correctly classified in order to increase the accuracy of predictions.

g) Overfitting in boosted decision trees is a major problem in high-dimensional data where the number of features is large compared to the number of samples.

2. What is model complexity and how does it play a role in the bias-variance tradeoff?

3. Imagine that you have a group of supervised algorithms which you train on different subsamples from the training set (i.e., by repeatedly drawing samples from the training dataset). Assuming you want to implement a bagging ensemble method, how would you pick a final predictor of a test data point if you have a …
a) classification task?
b) regression task?

4. Regarding random trees, why do we use randomness in selecting the features of each of the trees?

5. The Boosted Trees algorithm trains groups of small trees. What benefit does combining the prediction of all these trees bring?

Lecture 14: Stochastic Gradient Descent

1. Indicate whether the following is true or false. If something is false, indicate what is wrong with the statement justifying your response appropriately.

a) One particularity of the PEGASUS algorithm is that it computes the gradient for the optimization algorithm using only one sample out of the training data points – instead of using the whole dataset – thus increasing its computational efficiency.

b) PEGASUS has the disadvantage that the larger the training dataset, the slower it can be optimized to reach a particular test error.

c) Using a larger step size leads to a smoother training loss curve.

d) SGD is less likely to get stuck in local minima than full batch gradient descent.

2. Explain the tradeoffs when choosing the step size in gradient descent, especially in relation to optimum.

Lecture 15: MLPs and Backprop

1. Indicate whether the following is true or false. If something is false, indicate what is wrong with the statement justifying your response appropriately.

a) MLP and decision trees are related in that they sequentially apply non-linear functions to predict.

b) Linear activations cannot be used to reduce the range of the data values.

c) Sigmoid function transforms a continuous score to a probability value.

d) Sigmoid functions are differentiable and bounded, properties that made them popular in applied settings for a long time.

e) The backpropagation principle is a result of the chain rule of differentiation.

f) Since ReLU is nondifferentiable, they have the difficulty of not allowing larger gradient values to flow through the network.

g) Stacking layers in general increase the expressivity of an MLP.

h) The vanishing gradient problem is when several weights in an MLP go to zero.

i) An MLP are good predictors but are not able to learn features from the data.

j) An MLP has higher bias and lower variance than a perceptron.

2. What is a potential problem with only stacking linear layers in an MLP without non-linearities?

3. What is the problem with only using sigmoid functions in an MLP?

4. We saw in class that any function can be approximated to arbitrary accuracy by a network with two hidden layers with sigmoid activation. Someone might argue it wouldn't make sense to train an MLP with more than two layers. Provide an argument against such assertion.

Lecture 16: CNNs and Computer Vision

1. Indicate whether the following is true or false. If something is false, indicate what is wrong with the statement justifying your response appropriately.

a) Convolutional layers are needed to extract features and to reduce the dimensions of the input image data.

b) The weights of the convolutional layer can be trained.

c) Applying pooling to convolutional neural networks (CNN's) help reduce the feature representation.

d) An advantage of convolution is that it implements nonlinearity easily in a CNN.

e) Momentum used in SGD allows to skip all local minima.

f) Adaptive gradient (AdaGrad) was introduced to allow all weights to have an influence on the direction of the optimization trajectory.

g) AdaGrad provides smoother optimization trajectories than SGD, and thus can converge faster.

h) Adaptive Moment Estimation (Adam) fuses momentum with gradient normalization.

i) The residual connections of a ResNet alleviates the problem of vanishing gradients, though the neural networks cannot still be made too deep.

2. What is the reason why classic neural networks increase both their training and test error as their depth increases (for different training iterations)?

Lecture 17: Model Training and Tuning

1. Indicate whether the following is true or false. If something is false, indicate what is wrong with the statement justifying your response appropriately.

a) ResNets make connections between layers that improve the gradient flow during training.

b) Bottleneck ResNets reduce the dimensionality after performing convolution with the purpose of improving computational complexity of convolution during training and inference.

c) The idea with batch normalization is to normalize the features with their empirical mean and variance because feature distributions change "considerably" during training.

d) Batch normalization is not done over the whole dataset because this can adversely produce unexpected oscillations on the features learned in the hidden layers.

e) Padding is done over an image in a CNN to allow the convolution to be applied over the image's border.

f) The stride in a convolutional layer indicates the length of the convolution filter.

g) Data augmentation should generate the same type of features (or responses) at every layer of the deep network.

h) Linear probing can use a decoder to train a deep neural network over new classes, based on a previously trained feature extractor or encoder.

i) The “pre-compute features method” loads a pre-trained model which acts as a feature extractor, removes the last layer and replaces it with a new untrained layer, forming a “new” neural network over which to train the whole network.
j) Fine-tuning and linear probing in the “freeze encoder method” only differ in that the former has a smaller learning rate than the latter.
k) “Warm start” sets lower training rates for earlier layers during fine-tuning.

2. Why is fine-tuning okay when the training converges to a local minimum which is very close to the initialization point?

3. Order the methods of a) “fine-tuning”, b) “train from scratch”, c) “linear probe”, according to which one is better to use when there is a lot of new data to which one is better to use when there is scarce new data. Give a reason for the ordering.

Lecture 18: Words and Attention

1. Indicate whether the following is true or false. If something is false, indicate what is wrong with the statement justifying your response appropriately.

a) There is no single prescribed rule to represent a word as a token and it largely depends on the learning algorithm and the relationship it tries to capture between words.
b) An advantage of mapping words to integers is that it allows its effective use when training texts with limited vocabulary.
c) A problem with representing characters as integers is that representing texts can become very lengthy.
d) WordPiece Tokenizer represents most common subwords as tokens in an iterative way.
e) Compared to token-based approaches, Word2Vec needs to add additional dimensions of the vector representation when we need to add new words.
f) A disadvantage of the CBOW method for Word2Vec is that it cannot be trained using subgradient descent.
g) Both CBOW and Skip-Gram methods aim at increasing the likelihood of finding a word based on its surrounding words.
h) The Skip-Gram method trains Word2Vec so that a vector embedding of a word predicts well surrounding words (across a window over a text) using a linear model.
i) Consider the following sequence of four values (1,200,40,160). Then, using self-attention, with a similarity score of $S(x,y)=1/((x-y)^2+1)$, we can expect that after a theoretically infinite number of iterations, we will have the vector as (a,b,a,b), where a and b are two distinct numbers.
j) Self-attention can bring together values as a clustering method.
k) Transformers are only used in natural language processing because they only receive tokens as input.
l) Transformer blocks, which contain the multi-head attention processing, can be stacked to increase the processing power of the transformer.

m) Since transformers can process sequential data, they are very sensitive to the position that the input token has from the input stream of data.

2. Let the vectors p,q,r,s represent "Bill Gates", "Microsoft", "Tesla", "Jeff Bezos" in Word2Vec. Write the result you would expect for a properly trained Word2Vec on a very large body of text; if an answer cannot be known with certainty, just write "?".

a) q - p + s =
b) r - p + s =
c) p - q + r =
d) q - p + r =

Lecture 19: Transformers in Language and Vision

1. Indicate whether the following is true or false. If something is false, indicate what is wrong with the statement justifying your response appropriately.

a) Multiple-head attention allows the computing of different types of similarities.

b) The drop-out technique for transformers (and neural networks), which consists of ignoring some of the weights during training, has the purpose of accelerating (making more efficient) the training process.

c) An advantage of using CNNs for image (or video) processing over using Transformers is that they are more computationally efficient.

d) BERT is a bidirectional model in the sense that when it processes a word, we can set it to either consider the words that come later in the text, or instead consider the ones that come before.

e) One way in which BERT is trained is by masking tokens on sentences so that it is forced to predict the most likely token that should be there instead.

f) One way in which BERT is trained is by predicting whether, given two consecutive sentences, the second one "truly" succeeds the first one.

g) Though BERT is a powerful model, a difficulty in its use is that it needs to be completely retrained every time we want to use it in a different corpus.

h) The same transformer blocks developed for language processing can be used for vision processing (computer vision).

i) When training CNNs and ViT (Vision Transformers), both architectures allow the pixels on one part of the image to be equally compared to ones from even distant parts of the image.

j) Transformers could be used in applications where we need to synthesize a text which describes some input image.

k) Transformers could be used in applications where we need to synthesize an audio based on some input image.

l) Unified-IO is able to perform a wide range of vision-language tasks because the architecture does not include a task-specific decoder. Instead, there is an encoder and decoder for each modality, so text and image is translated into text and image. Many tasks fit into this framework, since the task definition itself can be represented in text.

Lecture 20: Foundation Models: CLIP and GPT-3

1. Indicate whether the following is true or false. If something is false, indicate what is wrong with the statement justifying your response appropriately.

a) Foundation models are based on the idea of pretraining a model in one task so that it can generalize to learn new unseen tasks.

b) ImageNet networks such as AlexNet are not an example of a foundation model because it only performs the task of classifying images it has seen in the ImageNet repository.

c) GPT1, like BERT, benefits from the bidirectionality of the sequence of tokens during training.

d) One innovation of GPT2, generally compared to classic supervised ML or even GPT1 and BERT, is that it conditions the output on the type of task at hand.

e) GPT2's architecture is fundamentally similar to GPT1 and BERT because it uses transformers.

f) Zero-shot learning is the ability of a learning model to correctly predict data from a distribution different from the one it was trained with.

g) Few-shot learning, e.g., like the one used in GPT3, consists in providing the description of the news task to the model along with an example of a desired input/output.

h) In the CLIP model, the encoders for text and images have been solely used for image prediction, including zero-shot prediction.

2. Why is GPT2's architecture in principle able to translate a phrase to Spanish or instead write a poem using such same phrase?

Lecture 21: Ethics and Impact of AI

1. List 3 different benefits and 3 different dangers of AI.

2. Suppose a self-driving car with 2 passengers (one in the driver's seat) crashes with another vehicle. List as many parties as possible that could be considered responsible for the crash. Then pick one of these parties that you consider most responsible and argue why.

3. Pick one of your favorite social media platforms (or choose any one if you do not have a favorite). Look through their Terms and Conditions or Settings and try to determine if you agreed to allow the company to use your data in training their ML models. Do you explicitly remember agreeing/forbidding the platform in using your data? Do you agree/disagree with how transparent/opaque the platform made finding this information? Write 1 paragraph on your findings.

Lecture 22: Bias in AI, Fair ML

1. Indicate whether the following is true or false. If something is false, indicate what is wrong with the statement justifying your response appropriately.

a) A type of bias in different corpora of texts is due to the mismatch between the frequency of words appearing in them and the "frequency" with which those same words are used in real life conversations/situations.

b) The only way the process of annotating unlabeled data can introduce bias is by "mislabeling".

c) The problem with confirmation bias is that people reject any type of information on the internet that does not correspond to their preconceived ideas.

d) The "bias network effect" is a bias that occurs when the output of some machine learning (ML) system outputs a prediction influenced by different biases along the ML pipeline (data gathering, hyperparameter tuning according to our own expectations, etc.), and in turn such prediction affects the future data that will be entered in the ML system again, creating a self-reinforcing bias loop in the system.

e) A sure way to guarantee the unbiasing of predictions in an ML system is by removing those "features" that may introduce some bias (e.g., when we suspect that certain attributes of the data may affect the prediction output due to cultural stereotypes).

f) Two ways to possibly mitigate biases during the training of an ML system is to let the test data cover edge cases, and use multiple distributions of data for training.

g) The idea of multi-task adversarial learning is to train features that are good at predicting the target labels, but bad at predicting sensitive attributes. This is done by sending negative gradients for the loss of predicting sensitive attributes through the feature backbone.

Lecture 25: Reinforcement Learning [SP'24]

1. Indicate whether the following is true or false. If something is false, indicate what is wrong with the statement justifying your response appropriately.

a) In reinforcement learning (RL), the current state of the environment does not depend on which states the agent has visited before the previous time-step, unless the history is also stored in the current state.

b) Deterministic policies are "deterministic" in name only because the underlying environment is a Markov decision process whose transition to different states – which depends on the action taken at each step – is stochastic in nature.

c) The objective of an RL agent is to maximize its expected cumulative reward.

d) In deep RL, the main purpose of deep neural networks is to reduce the dimensionality of the input data that defines the state of the environment – e.g. in robotics, the images captured by a camera define the states of the environment.

e) The Bellman equation is used to characterize the optimal policy and is the basis of Q-learning.

f) Approximate Q-learning is used when the state space is continuous, e.g., in robotic control applications.
g) The problem with sparse rewards (rewards that are only non-zero in perhaps one or relatively very few states of the environment) is that it takes more time for the agent to realize whether it has been taking good actions.
h) In RL, it is always inconvenient to incentivize exploration after the initial stages of running the learning algorithm.

2. Imagine you want a robot to learn how to go from point A to point B using RL. Assume the reward function has been carefully designed and can't be modified. After training, it seems the robot takes longer than expected to do the task. What could you try to do in order to decrease the amount of time the robot will take to do the task? Why?

3. Increasing exploration in an RL algorithm will generally allow the agent to know more about the environment that surrounds it — and knowing more about the environment logically provides more information for coming up with a better policy to navigate it. How could increasing exploration by a large margin be possibly detrimental to an RL algorithm?

Lecture 26: Audio and 1-D Signals [SP'24]

1. Indicate whether the following is true or false. If something is false, indicate what is wrong with the statement justifying your response appropriately.

a) Fourier series only work well for signals that are periodic.
b) Fourier series can be expressed as a weighted sum of sines and cosines.
c) A square wave of period T has very strong components in very high frequencies much larger than 1/T.
e) Only knowing the amplitudes and frequencies of all the sinusoids comprising a specific sound is enough to reconstruct it.
f) Since audio is a 1D signal, we cannot use traditional neural network architectures designed for image classification – recall that an image can be seen as a 2D signal.
g) Any ASR solution (Audio to Speech Recognition) trained with one person will have no difficulty in recognizing the characters of the words spoken by another person.
h) The filter [1/9 2/9 1/3 2/9 1/9] is used for smoothing 1D signals.
i) In forecasting methods for 1D signals (e.g., using neural networks), the length of the window that is chosen as past history of the current input (used for forecasting) is a hyperparameter.