2018-03-12

Machine learning is often divided into the categories Supervised, Unsupervised and Reinforcement learning. Categorize the following three learning methods accordingly:
Q-Learning: Reinforcement learning
k-means: Unsupervised learning
kNN: Supervised learning

Mention a classifier that uses the maximum margin principle
Support vector machines

What is the basic requirement on an activation function for being used in a multi layer perceptron with back propagation?
Differentiable

You have a high dimensional data set where many of the features are correlated. Suggest a proper pre-processing before feeding that data into a classifier:
PCA

Explain briefly the exploration exploitation dilemma in reinforcement learning.
The dilemma refers to the conflict between utilizing the current knowledge to maximize the reward and trying new policies to explore new potentially better solutions.

We have 10 samples in a 20-dimensional space that we want to analyze with a kernel method. How large is the kernel matrix?
10x10

What is being optimized by Linear Discriminant Analysis (LDA)?
The quotient between the difference between the cluster means and the variance of the projections of the clusters.

Mention a method that can speed up gradient search?
Using a momentum term

Explain or draw an example of when a sigmoid function in the output layer of a linear classifier trained by minimizing the means square error can improve the accuracy.
It will improve the accuracy when the distribution of the classes are very different, since the boundary will be in the middle between the class centroids.

What is the benefit of a convolutional layer compared to a fully connected layer in a multi-layer neural network?
One benefit is that there are fewer parameters to train, which reduces the risk of over-fitting and improves the generalization properties.

2019-03-18

Which of the following methods are supervised learning methods?
Mixture of gaussians: no
Back-propagation: yes
kNN: yes
LDA: yes
SVN:yes
PCA: no

Explain the purpose with the so-called slack variables in Support Vector Machines:
Slack variables are used to allow some samples to be miss-classified, in order not to verfit to noisy data and outliers.

Which of these functions can be used in the hidden layers of a back-prop network?
y = s : no, it is linear
y = tanh(s) : yes
y = s/||s|| : no, not differentiable
y = e^(-s**2) :yes

What is described by the first eigenvalue of the data covariance matrix?
The maximum variance of the data

What problem can be illustrated by the multi-armed bandit?
The multi-armed bandit illustrates the exploration-exploitation problem

Write the general definition of a kernel function k(x,y)
k(x,y) = phi(x)**T*phi(y) where phi is a non-linear function

k-means clustering and mixture of gaussians are two examples of a general optimization method. What is this method called?
The EM-algorithm

What do you get if you combine Bagging and decision trees?
A random forest

What is described by the empirical risk?
The empirical risk is the number of wrong classifiers

What is the purpose of the hidden layers in a multi-layer perceptron classifier?
The purpose of the hidden layers in a multi-layer perceptron classifier is to transform the data to a space where the problem is linearly separable

2019-06-12

Classify the following learning methods as Supervised or unsupervised:
kNN: S
SVM: S
AdaBoost: S
PCA: U
Multi-layer perceptron: S
Mixture of gaussians: U

How is the accuracy of a classifier calculated?
By the number of correctly classified samples divided by the total amount of samples

Why are the following two functions not useful in the hidden layers in the back-propagation neural network?
y = s, it is linear
y = sign(s), it is not differentiable

What is described by the first eigenvector of the data covariance matrix?
It shows the direction in which the data has maximum variance

What assumption is made about the distributions of the two classes in linear discriminant analysis?
That both classes have the same covariance matrix, i.e. that the shapes of the distributions are identical

Suppose that you know the Q-function values for a certain state. How do you determine the V-value for that state?
Max(Q(s,a)), the value of the V-function is the maximum Q-value over all possible actions

What is the purpose with a momentum term in gradient descent?
To improve speed of convergence

In which kind of learning tasks are linear units more useful than sigmoid activation functions in the output layer of a multi-layer neural network?
In regression tasks, in particular when the output needs to be outside the interval -1 to 1

Explain the purpose with the so-called slack variables in the Support Vector Machines.
Slack-variables are used to allow some samples to be miss-classified, in order not to overfit to noisy data and outliers.

All the weights of one layer in a neural network can be described as a matrix W. Describe an important property of this matrix for a convolutional layer in a CNN.
The weight matrix in a convolutional layer is sparse (and with constant diagonals i.e. a Toeplitz matrix)

2019-08-31

Machine learning can be divided into supervised, unsupervised and reinforcement learning. Mention one example of learning methods for each of these three classes:
S: Back-propagation neural networks
U: Mixture of gaussians
R: Q-learning

What happens if the input to the bias weight in a perceptron is set to the constant value 2?
The input value to the bias weight does not matter, as long as it is not zero

Connect the concepts
Maximum margin principle - Support vector machine
Back propagation - Neural Network
Brute force optimization - Decision tree

What is the difference between k-means clustering and mixture of gaussians clustering?
In k-means clustering only the centers of the clusters are modelled, whereas in mixture of gaussians clustering, also the cluster shape is modelled using the covariance matrix, which can model circular and elliptical cluster shapes.

In Linear Discriminant Analysis, a quotient between a distance d and a variance v is maximized. What is described by d and v?
d is the distance between the cluster centers and v is the variance of the projected clusters.

Explain the principle of e-greedy exploration in the context of reinforcement learning.
Having an e chance of exploring in another direction to combat the exploitation-exploration dilemma.

The following update rule is sometimes used in gradient search. What is the term called?
Momentum term.

In the figure below, a simple one-layer perceptron has been used for classification. Give an explanation in terms of the activation function for the different resulting discriminant functions. (Solid vs dashed)
A linear function would give the result according to the dashed line while a sigmoid activation function would give a result similar to the solid line

What output will you get if you use one of the support vectors as input to a support vector machine
+-1

Mention a type of neural network that uses weight sharing(parameter sharing)
CNN

2020-03-20

Which of the following methods are unsupervised learning methods?
Mixture of gaussians : y
Back-propagation: n
kNN: n
k-means: y
SVM: n
PCA: y

What output will you get if you use one of the support vectors as input to a support vector machine?
+-1

What are the two basic requirements on a function for being used as an activation function in the hidden layers of a multi-layer perceptron with back propagation?
Nonlinear and differentiable

Draw the second principal components of the distribution in the figure below!
See image

We want to train a perceptron to separate the two classes in the figure below. How many parameters wi do we need to optimize?
There are three parameters to train. Two input weights and one bias weight.

Draw a plausible separating line and mark the support vectors for a linear SVM in the figure below.
See image

Assume you have a set of data points in a three-dimensional feature space and you want to cluster the points into two clusters using Mixture of gaussians. How many scalar numbers need to to be estimated in the optimization, disregarding the set membership variables Si?
We need two mean vectors(each 3 parameters) and two covariance matrices(9 parameters each) which gives 6+18=24 parameters in total. Since the covariance matrices are symmetric we only need to estimate 6 parameters for each covariance matrix which gives 18 total.

Consider the following non-linear mapping from the input space to a higher-dimensional feature space:
rho(x) = (sin(x), cos(x))
Write the corresponding kernel function k(x,y)
k(x,y) = sin(x)*sin(y)+cos(x)*cox(y)

How do you measure the generalization error?
Test a model on previously unseen data.

How should you change the value of k in k-NN if you want to decrease the risk of overfitting?
The value of k should increase.

Suggest two kernel functions that can be used to separate the two classes in the figure below by applying the kernel to a linear classifier
A gaussian or quadratic kernel would be appropriate

Perform two iterations with k-means in the figure below.
It has converged, see image

Are the two classes in the figure below linearly separable? How will LDA perform on this task?
Yes they are linearly separable. But linear discriminant analysis will perform poorly on this task since the two classes have very different distributions and LDA assumes similar distributions of the classes.

The table below shows the Q-values for different states Si and actions Ai. What are the value function values V(Si) for all states and what action will the system take in state 2 if i follows a greedy policy?
V(S1) = 3
V(S2) = 5
V(S3) = 7
V(S4) = 4
It will take action 2

2020-06-10

Which of the following methods are supervised learning methods?
PCA: f
Back-propagation: t
k-means: f
kNN: t
LDA: t
SVN: t

Write the cost function that is being optimized in Support Vector Machines
min||w||**2 under the constraint yi(w**Txi + w0) <= 1

Which of these functions can be used in the hidden layers of a back prop network?
1,3,4

Draw the first principal component of the distribution in the figure below!
See image

We want to train a perceptron to separate the two classes in the figure below. How many parameters wi do we need to optimize?
There are two parameters to train. One input weight and one bias weight.

Assume you have a set of data points in a two-dimensional feature space and you want to cluster the points into three clusters using mixture of gaussians How many scalar numbers need to be estimated in the optimization, disregarding the set membership variables Si?
We need three mean vectors(each 2 parameters) and three covariance matrices(4 parameters each) which gives 6+12=18 parameters in total. Since covariance matrices are symmetric we actually only need to estimate 3 parameters for each covariance matrix, which gives 15 total

Mention an alternative to the sigmoid activation function that can be used to avoid the vanishing gradient problem in deep neural networks.
The rectified linear unit(ReLU) is used to avoid the vanishing gradient problem in deep multi-layer networks.

Why isn’t it optimal to try to reach the lowest possible error on the training data when training a neural network?
This leads to overfitting

What is determined by the parameter k in kNN
How many points that vote on the classification.

Why isn’t it always best to choose the response with the best Q-value in Q-learning?
In order to have the possibility to find an even better policy(the bias-variance dilemma)

Consider the following non-linear mapping of the input data x:
You want to analyze this data with a kernel method. How is the scalar product phi(x)**Tphi(y) expressed in the input data space?
x1**2*y1**2 + x2**2*y2**2 + 2x1x2y1y2 = (x1y1 + x2y2)**2 = (x**T*y)**2

Perform two iterations with k-means in the figure below. The dots indicate the data points and the crosses are the two prototypes(k=2). Show both steps. Has the algorithm converged after these steps?
It has converged after only 1 step

Assume that a one-layer neurla network is trainged on the data in the figure below. Draw the approximate zero-crossing of the resulting discriminant function if a) a linear activation is used and b) a sigmoid activation function is used.
See figure

Draw a decision tree that implements the follow discriminant function
See image

The table below shows the Q-values for different states Si and actions Ai. What are the values V(Si) for all states and what action will the system take in state 2 if it follows a greedy policy?
V(S1) = 4
V(S2) = 4
V(S3) = 3
V(S4) = 2
It would take action 1

2020-08-29

Machine learning is often divided into the categories supervised, unsupervised and reinforcement learning. Categorize the following three learning methods accordingly:
Q-learning: Reuinforcement
k-means: Unsupervised
kNN: Supervised

Mention a classifier that uses the maximum margin principle
SVM

Which of these functions can be used in the hidden layers of a back-prop network?
2,4

What can be said approximately about the two eigenvalues of the data covariance matrix for the data points in the distribution below?
They are approximately equal

We want to train a single layer perceptron to separate objects into two classes. A pre-processing step measures the height and width of each object and the classification should be based on these features. How many parameters do we have to optimize when we train the perceptron?
3, 2 input weights and one bias weight.

Assume you have 100 data points in a one dimensional feature space and you want to cluster the points into three clusters using a mixture of gaussians. How many scalar numbers need to be estimated in the optimization, disregarding the set membership variables Si?
A three mean vectors (1 param each) and three covariance matrices (1 param each) 3+3=6

What is defined by a kernel function?
The kernel function defines a scalar product between the vectors mapped to a feature space.

How can you notice if a supervised machine learning algorithm has overtrained?
It performs much better on the training data than unseen data.

Mention a classifier that does not use a parameterized discriminant function.
kNN

If you have 100 training examples and perform a 4 fold cross validation to evaluate the performance of a classifier, how many times must you train the classifier?
4 times

Consider a polynomial kernel function k(x1,x2) = 1 + (x1**T*x2)**2, what is the distnace between two feature vectors x1 = (1,1) and x2 = (0,1) in the new feature space defined by this kernel function? fuck you, see image

Perform two iterations with k-means in the figure below. The dots indicate the data points and the crosses two prototypes(k=2). Show both steps. Has the algorithm converged? no

We want to train a simple classifier on the data in the figure below. Draw the approximate zero-crossing of the resulting discriminant function if a) a linear perceptron is used and b) a linear SVM is used
See image

ReLU activation functions are more and more used in neural networks instead of the tanh activation function. Draw both activation functions and give a) an advantage of the ReLU function compared to the tanh function. b) a disadvantage of the ReLU function compared to the tanh function.

a) No vanishing gradient problem
b) Knockout problem, a neuron can be driven to a state where it never activates for any input

Draw decision tree
See image

2021-03-19

Describe the difference in terms of training data between supervised and unsupervised learning.
Supervised is labeled.

Draw the line for which the discriminant function f(x) = 0 for a linear SVN without slack variables and mark the support vectors in the figure below.
See image

Draw the line for which the discriminant function f(x) = 0 for a single perceptron with a linear activation function and also for a sigmoid activation function.
See image

Draw the first principal component of the distribution in the figure below
See image

Why is the function f(s) = s/|s| not useful as an activation function in a multi-layer back propagation network?
It has a zero derivative everywhere except for 0 where it is undefined

What are networks called where some layers are trained to learn f(x)-x?
Residual networks

The following update rule is sometimes used in gradient search, what is it called?
Momentum term

What has happened when the validation error is much larger than the training error?
Overfitting

Consider the following non-linear mapping from the N-dimensional input space to a non-linear feature space. Write the corresponding kernel function k(x,y)
k(x,y) = sum i = 1..N(e**-(xi**2+y**2))

Which class does the data sample x belong to using a kNN classifier with k = 1 and k = 3?
1: circle
3: square

Write a 3-by-3 convultion kernel that detects horizontal lines in an image. The kernel should not be sensitive to the mean intensity of the image neighborhood.
-1 -1 -1
2 2 2
-1 -1 -1

Draw the approximate cluster centers and the border between the two classes after successful convergence for 1) a k-means and 2) mixture of gaussians assuming 2 classes in both cases. Number of points in distributions are equal. See image

Assume you know that the optimal discriminant function for a particular classification problem can be defined as...
Suggest another discriminant function that can be used if you want to adapt the function to you data using gradient descent and write the update equation for the parameters in that function.
See image

Consider the two different network architectures below. The height of each box indicates the size of each layer. Which one could be used for an image classification task and which one could be used for image segmentation?

a) can be used for image segmentation as it has the same size of the output layer as the input layer
b) can be used for image classification as the output is of lower dimensionality than the input

15. Fuck you.