Intro to ML Quiz

* Required

Email address *

Your email

Which of the following statements are true? Select all that apply. 1 point

- [ ] K-Means will always give the same results regardless of the initialization of the centroids.
- [ ] A good way to initialize K-means is to select K (distinct) examples from the training set and set the cluster centroids equal to these selected examples
- [ ] Once an example has been assigned to a particular centroid, it will never be reassigned to another different centroid
- [ ] If we are worried about K-means getting stuck in bad local optima, one way to ameliorate (reduce) this problem is if we try using multiple random initializations.

Logistic regression: For logistic regression gradient descent always finds the optimal solution given enough time 1 point

- ○ True
- ○ False

K-nearest neighbors (KNN): K-nearest neighbors (KNN) is a supervised learning algorithm which can be used for both classification and regression.

1 point

- True
- False

Match the most appropriate thing for the following. (Match the most specific one than generalised one)

4 points

	regression	Classification	Supervised Learning	unsupervised learning	reinforcement learning	Sem Superv Learn
the average of the K nearest neighbours in the datasets	○	○	○	○	○	○
a fixed number of parameters to learn a function	○	○	○	○	○	○
the number of parameters grows with the size of training data	○	○	○	○	○	○
Support Vector Machine (SVM)	○	○	○	○	○	○

Regression: The gradient for logistic regression is exactly the same as for linear regression. 1 point

- True
- False

Curse of dimensionality: A low dimensional sphere (or any other regular shape), most of the volume is close to the surface Whereas in a high dimensional sphere, most of the volume is close to the centre. 1 point

- True
- False

Match the following.

4 points

	determines the size of the steps required to reach a minimum cost.	finds the best model out of the many possible models	makes comparison between data sets easier and the results more meaningful to when performing regressions in different problem spaces.	results the regression with more uniform residuals and less drastic outliers.
Squaring residues in linear regression in cost function of linear regression	○	○	○	○
The advantage of taking mean of squares of residues in in cost function of linear regression	○	○	○	○
The Gradient Descent in linear regression	○	○	○	○
Learning rate in linear regression algorithm	○	○	○	○

Suppose a massive dataset is available for training a learning algorithm. Training on a lot of data is likely to give good performance when two of the following conditions hold true. Which are the two?

1 point

- [ ] Our learning algorithm is able to represent fairly complex functions (for example, if we train a neural network or other model with a large number of parameters).
- [ ] When we are willing to include high order polynomial features of x (such as x1^2, x2^2, x1x2, etc.).
- [ ] The classes are not too skewed.
- [ ] The features x contain sufficient information to predict y accurately. (For example, one way to verify this is if a human expert on the domain can confidently predict y when given only x).

Which of the following are true of collaborative filtering systems? Check all that apply.

1 point

- [ ] If you have a dataset of user ratings on some products, you can uses these to predict one user's preferences on products he has not rated.
- [ ] To use collaborative filtering, you need to manually design a feature vector for every item (e.g., movie) in your dataset, that describes that item's most important properties.
- [ ] For collaborative filtering, it is possible to use one of the advanced optimization algoirthms (L-BFGS/conjugate gradient/etc.) to solve for both the x(i)'s and θ(j)'s simultaneously.
- [ ] For collaborative filtering, the optimization algorithm you should use is gradient descent. In particular, you cannot use more advanced optimization algorithms (L-BFGS/conjugate gradient/etc.) for collaborative filtering, since you have to solve for both the x(i)'s and θ(j)'s simultaneously.

Suppose you have trained a logistic regression classifier which is outputting h(x). Currently, you predict 1 if h(x)>threshold, and predict 0 if h(x) <threshold, where currently the threshold is set to 0.5. Suppose you decrease the threshold to 0.1. Which of the following are true? Check all that apply. 1 point

- [ ] The classifier is likely to have unchanged precision and recall, and thus the same F1 score.
- [ ] The classifier is likely to now have higher recall.
- [ ] The classifier is likely to now have higher precision.
- [ ] The classifier is likely to have unchanged precision and recall, but lower accuracy.

Match the most appropriate one for the following. 2 points

	convex in shape	probability
output of the logistic function	○	○
cost function	○	○

Which of the following is not used to best k for K-nearest neighbors (KNN)? 1 point

- ○ cross-validation
- ○ k equal to the square root of total number of data points
- ○ log n where n is the total number of data points.

Dimensionality: Distance measurements becomes easy when different dimensions are correlated with each other, a common property of high dimensional datasets

1 point

- True
- False

Match the following

5 points

	L1 regularisation	L2 regularisation
penalty proportional to the squared magnitude of each weight	○	○
penalises the larger weights more	○	○
promotes a larger number of parameters w to be zero	○	○
smaller weights are equally penalized as larger weights	○	○
reduce the size of a model	○	○

Match the most appropriate thing for the following. (Match the most specific one than generalised one)

4 points

	ement ing	Semi-Supervised Learning	K-nearest neighbors (KNN)	K-nearest neighbors(KNN) regression algorithm	parametric model	non-parametric model	Max-margi classifi
A robot is in a maze, and it needs to find a way out	◯	◯	◯	◯	◯	◯	◯
Training an AI for a complex game	◯	◯	◯	◯	◯	◯	◯
some data that is labelled, and some that is not labelled	◯	◯	◯	◯	◯	◯	◯
dataset is the model	◯	◯	◯	◯	◯	◯	◯

Which of the following statements are true? Check all that apply. 1 point

- [ ] After training a logistic regression classifier, you must use 0.5 as your threshold for predicting whether an example is positive or negative.
- [ ] On skewed datasets (e.g., when there are more positive examples than negative examples), accuracy is not a good measure of performance and you should instead use F1 score based on the precision and recall.
- [ ] The "error analysis" process of manually examining the examples which your algorithm got wrong can help suggest what are good steps to take (e.g., developing new features) to improve your algorithm's performance.
- [ ] It is a good idea to spend a lot of time collecting a large amount of data before building your first version of a learning algorithm.
- [ ] If your model is underfitting the training set, then obtaining more data is likely to help.

Model accuracy and performance: High classification accuracy always indicates a good classifier. 1 point

- True
- False

The amount of rain that falls in a day is usually measured in either millimeters (mm) or inches. Suppose you use a learning algorithm to predict how much rain will fall tomorrow. Would you treat this as a classification or a regression problem? 1 point

- Classification
- Regression

Suppose you are working on a spam classifier, where spam emails are positive examples (y=1) and non-spam emails are negative examples (y=0). You have a training set of emails in which 99% of the emails are non-spam and the other 1% is spam. Which of the following statements are true? Check all that apply.

1 point

- [ ] If you always predict spam (output y\=1), your classifier will have a recall of 0% and precision of 99%.
- [ ] If you always predict non-spam (output y\=0), your classifier will have a recall of 0%.
- [ ] If you always predict spam (output y\=1), your classifier will have a recall of 100% and precision of 1%.
- [x] If you always predict non-spam (output y\=0), your classifier will have 99% accuracy on the training set, and it will likely perform similarly on the cross validation set.

Croos Validation: k-fold cross validation increases the computational requirements for training the model by a factor of k.

1 point

- ○ True
- ○ False

Match the most appropriate thing for the following. (Match the most specific one than generalised one)

4 points

	regression	Classification	Supervised Learning	unsupervised learning	reinforcement learning	Sem Superv Learn
Anomaly detection	○	○	○	○	○	○
Association	○	○	○	○	○	○
Netflix Data\: Given a user and a movie, predict the rating the user is going to give to the movie	○	○	○	○	○	○
the input itself depends on actions we take	○	○	○	○	○	○

You are training a classification model with logistic regression. Which of the following statements are true? Check all that apply.

1 point

- ☐ Introducing regularization to the model always results in equal or better performance on the training set.
- ☐ Adding many new features to the model makes it more likely to overfit the training set.
- ☐ Adding many new features to the model helps prevent overfitting on the training set.
- ☐ Adding a new feature to the model always results in equal or better performance on examples not in the training set.

Which of the following are true? Check all that apply. 1 point

- [ ] In anomaly detection, we fit a model p(x) to a set of negative (y\=0) examples, without using any positive examples we may have collected of previously observed anomalies
- [ ] If you are developing an anomaly detection system, there is no way to make use of labeled data to improve your system
- [ ] When choosing features for an anomaly detection system, it is a good idea to look for features that take on unusually large or small values for (mainly the) anomalous examples
- [ ] If you have a large labeled training set with many positive examples and many negative examples, the anomaly detection algorithm will likely perform just as well as a supervised learning algorithm such as an SVM.

In K-means is an iterative algorithm, and two of the following steps are repeatedly carried out in its inner-loop. Which two? 1 point

- [ ] Using the elbow method to choose K.
- [ ] The cluster assignment step, where the parameters c(i) are updated.
- [ ] The cluster centroid assignment step, where each cluster centroid μi is assigned (by setting c(i)) to the closest training example x(i).
- [ ] Move the cluster centroids, where the centroids μk are updated.

Suppose you have implemented regularized logistic regression to classify what object is in an image (i.e., to do object recognition). However, when you test your hypothesis on a new set of images, you find that it makes unacceptably large errors with its predictions on the new images. However, your hypothesis performs well (has low error) on the training set. Which of the following are promising steps to take? Check all that apply. 1 point

- [ ] Try decreasing the regularization parameter ?
- [ ] Get more training examples.
- [ ] Try adding polynomial features.
- [ ] Try increasing the regularization parameter

Suppose you have implemented regularized logistic regression to predict what items customers will purchase on a web shopping site. However, when you test your hypothesis on a new set of customers, you find that it makes unacceptably large errors in its predictions. Furthermore, the hypothesis performs poorly on the training set. Which of the following might be promising steps to take? Check all that apply.

1 point

- [ ] Use fewer training examples.
- [ ] Try to obtain and use additional features.
- [ ] Try using a smaller set of features.
- [ ] Try decreasing the regularization parameter

Match the most appropriate thing for the following. (Match the most specific one than generalised one)

5 points

	regression	Classification	Supervised Learning	unsupervised learning	reinforcement learning	Semi-Supervise Learning
maps inputs to desired outputs	○	○	○	○	○	○
output variable is a category	○	○	○	○	○	○
output variable is a real value	○	○	○	○	○	○
identify patterns in the data	○	○	○	○	○	○
discover groups of similar users	○	○	○	○	○	○

Match the following with appropriate methods of recommendation. 3 points

	collaborative filtering by matrix factorization method	Content-based recommendation	Collaborative filtering
systems rely solely on information about the user	○	○	○
similarity between users and objects	○	○	○
cosine distance between users	○	○	○

Which of the following statements are true? Check all that apply. 1 point

- ☐ The performance of a learning algorithm on the training set will typically be better than its performance on the test set.
- ☐ Suppose you are training a logistic regression classifier using polynomial features and want to select what degree polynomial (denoted d in the lecture videos) to use. After training the classifier on the entire training set, you decide to use a subset of the training examples as a validation set. This will work just as well as having a validation set that is separate (disjoint) from the training set.

Multi class regression: In multi-class logistic regression, the model parameters W is a matrix of size k X n where n is the number of classes and k is the number of input features. 1 point

- ○ True
- ○ False

You run a movie empire, and want to build a movie recommendation system based on collaborative filtering. There were three popular review websites (which we'll call A, B and C) which users to go to rate movies, and you have just acquired all three companies that run these websites. You'd like to merge the three companies' datasets together to build a single/unified system. On website A, users rank a movie as having 1 through 5 stars. On website B, users rank on a scale of 1 - 10, and decimal values (e.g., 7.5) are allowed. On website C, the ratings are from 1 to 100. You also have enough information to identify users/movies on one website with users/movies on a different website. Which of the following statement is true?

1 point

- It is not possible to combine these websites' data. You must build three separate recommendation systems
- You can combine all three training sets into one as long as your perform mean normalization and feature scaling after you merge the data
- You can combine all three training sets into one without any modification and expect high performance from a recommendation system
- You can merge the three datasets into one, but you should first normalize each dataset's ratings (say rescale each dataset's ratings to a 0-1 range).

Match the following for linear regression.

4 points

	Small cost	High cost
Higher residuals	○	○
Small residuals	○	○
line is far from all the points	○	○
line is close to the points	○	○

SVM: For SVM model, minimizing loss is equivalent to maximizing-margin. 1 point

- ◯ True
- ◯ False

Match the following 3 points

	overfitting	Regularisation	simpler model
Penalise large weights	◯	◯	◯
weights large	◯	◯	◯
weights small	◯	◯	◯

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 might be promising steps to take? Check all that apply 1 point

- ☐ Try using a neural network with a large number of hidden units.
- ☐ Reduce the number of examples in the training set.
- ☐ Increase the regularization parameter λ.
- ☐ Use an SVM with a Gaussian Kernel

Match the following with respect to K-means clustering

3 points

	Rule of thumb	Elbow method	Silhouette analysis
K ≈ sqrt(N/2)	○	○	○
select k for which k "the total within-cluster sums-of-squares (WCSS)" is optimum	○	○	○
separation distance between clusters	○	○	○

Which of the following are recommended applications of PCA? Select all that apply.

1 point

- [ ] To get more features to feed into a learning algorithm.
- [ ] Data compression: Reduce the dimension of your input data x(i), which will be used in a supervised learning algorithm (i.e., use PCA so that your supervised learning algorithm runs faster).
- [ ] Data visualization: Reduce data to 2D (or 3D) so that it can be plotted.
- [ ] As a replacement for (or alternative to) linear regression: For most learning applications, PCA and linear regression give substantially similar results.

You want to use a learning ...::Suppose you are working on stock market prediction, and you would like to predict whether or not a particular stock's price will be higher tomorrow than it is today. You want to use a learning algorithm for this. Would you treat this as a classification or a regression problem?

1 point

○ Classification

○ Regression

For a regularisation parameter (usually denoted as C), match the following. 3 points

	larger margin but misclassify more points	small margin but classify more points correctly
large values of C	○	○
small values of C	○	○
more outliers	○	○

Which of the following statements are true? Check all that apply. 1 point

- [ ] Gradient checking is useful if we are using gradient descent as our optimization algorithm. However, it serves little purpose if we are using one of the advanced optimization methods (such as in fminunc).
- [ ] If our neural network overfits the training set, one reasonable step to take is to increase the regularization parameter
- [ ] Using gradient checking can help verify if one's implementation of backpropagation is bug-free.
- [ ] Gradient checking is useful if we are using one of the advanced optimization methods (such as in fminunc) as our optimization algorithm. However, it serves little purpose if we are using gradient descent.

Suppose you have an unlabeled dataset {x(1),...,x(m)}. You run K-means with 50 different random initializations, and obtain 50 different clusterings of the data. What is the recommended way for choosing which one of these 50 clusterings to use? 1 point

- ○ Always pick the final (50th) clustering found, since by that time it is more likely to have converged to a good solution.
- ○ For each of the clusterings, compute 1/m Σi\=1 to m ||x(i) − μc(i) ||2, and pick the one that minimizes this
- ○ Use the elbow method.
- ○ Manually examine the clusterings, and pick the best one.

Suppose you have a dataset with m=1000000 examples and n=200000 features for each example. You want to use multivariate linear regression to fit the parameters to our data. Should you prefer gradient descent or the normal equation? 1 point

- ○ The normal equation, since gradient descent might be unable to find the optimal
- ○ Gradient descent, since it will always converge to the optimal
- ○ The normal equation, since it provides an efficient way to directly find the solution.
- ○ Gradient descent, since (XTX)power of -1 will be very slow to compute in the normal equation

Which of the following statements are true? Check all that apply 1 point

- ☐ PCA is susceptible to local optima; trying multiple random initializations may help.
- ☐ Even if all the input features are on very similar scales, we should still perform mean normalization (so that each feature has zero mean) before running PCA.
- ☐ Given an input x∈Rn, PCA compresses it to a lower-dimensional vector z∈Rk.
- ☐ PCA can be used only to reduce the dimensionality of data by 1 (such as 3D to 2D, or 2D to 1D).

Cost function:Non-convex cost function, does gradient descent get stuck on local minima. 1 point

- ○ True
- ○ False

In smaller dataset, it is much tougher to separate reliable patterns from noise and leads to a good generalisation. 1 point

- ☐ True
- ☐ False

Which of the following statements are true? Check all that apply. 1 point

- ☐ If a neural network is overfitting the data, one solution would be to increase the regularization parameter
- ☐ The activation values of the hidden units in a neural network, with the sigmoid activation function applied at every layer, are always in the range (0, 1).
- ☐ A two layer (one input layer, one output layer; no hidden layer) neural network can represent the XOR function.

Which of the following are reasons for using feature scaling? 1 point

- ○ It speeds up gradient descent by making each iteration of gradient descent less expensive to compute.
- ○ It speeds up gradient descent by making it require fewer iterations to get to a good solution.
- ○ It is necessary to prevent gradient descent from getting stuck in local optima.
- ○ It is necessary to prevent the normal equation from getting stuck in local optima.

Match the following

3 points

	performs well on training data but does not perform well on test data	The ability to perform well on unseen data	does not perform well on training data as well as unseen data
Generalisation	◯	◯	◯
over-fit	◯	◯	◯
under-fit	◯	◯	◯

Suppose you have trained an anomaly detection system for fraud detection, and your system that flags anomalies when p(x) is less than ε, and you find on the cross-validation set that it is missing many fradulent transactions (i.e., failing to flag them as anomalies). What should you do?

1 point

- ◯ Decrease ε
- ◯ Increase ε

A computer program is said to learn from experience E with respect to some task T and some performance measure P if its performance on T, as measured by P, improves with experience E. Suppose we feed a learning algorithm a lot of historical weather data, and have it learn to predict weather. In this setting, what is E?

1 point

- ◯ The weather prediction task
- ◯ The process of the algorithm examining a large amount of historical weather data.
- ◯ The probability of it correctly predicting a future date's weather.

Curse of dimensionality: On high dimensional datasets, the concept of nearest neighbour isn't as effective as on low dimensional datasets. 1 point

- ◯ True
- ◯ False

High values of the learning rate hyper-parameter are always preferred to reach minimum very fast. 1 point

- ◯ True
- ◯ False

Which of the following statements are true? Check all that apply. 1 point

- ☐ If the training and test errors are about the same, adding more features will not help improve the results.
- ☐ When debugging learning algorithms, it is useful to plot a learning curve to understand if there is a high bias or high variance problem.
- ☐ A model with more parameters is more prone to overfitting and typically has higher variance.
- ☐ If a learning algorithm is suffering from high bias, only adding more training examples may not improve the test error significantly.

In which of the following situations will a collaborative filtering system be the most appropriate learning algorithm (compared to linear or logistic regression)? 1 point

- [ ] You manage an online bookstore and you have the book ratings from many users. You want to learn to predict the expected sales volume (number of books sold) as a function of the average rating of a book
- [ ] You run an online bookstore and collect the ratings of many users. You want to use this to identify what books are "similar" to each other (i.e., if one user likes a certain book, what are other books that she might also like?)
- [ ] You've written a piece of software that has downloaded news articles from many news websites. In your system, you also keep track of which articles you personally like vs. dislike, and the system also stores away features of these articles (e.g., word counts, name of author). Using this information, you want to build a system to try to find additional new articles that you personally will like.
- [ ] You manage an online bookstore and you have the book ratings from many users. For each user, you want to recommend other books she will enjoy, based on her own ratings and the ratings of other users.

Model performance on train and test datasets: The model that gives the lowest error on training data will also give the lowest error on test data. 1 point

- ○ True
- ○ False

For which of the following tasks might K-means clustering be a suitable algorithm? Select all that apply. 1 point

- [ ] From the user usage patterns on a website, figure out what different groups of users exist.
- [ ] Given historical weather records, predict if tomorrow's weather will be sunny or rainy.
- [ ] Given many emails, you want to determine if they are Spam or Non-Spam emails.
- [ ] Given a database of information about your users, automatically group them into different market segments.

Which of the following statements are true? Check all that apply 1 point

- [ ] It is important to perform feature normalization before using the Gaussian kernel.
- [ ] If the data are linearly separable, an SVM using a linear kernel will return the same parameters θ regardless of the chosen value of C (i.e., the resulting value of θ does not depend on C).
- [ ] The maximum value of the Gaussian kernel (i.e., sim(x,l(1))) is 1.
- [ ] Suppose you are using SVMs to do multi-class classification and would like to use the one-vs-all approach. If you have K different classes, you will train K - 1 different SVMs

Which of these is a reasonable definition of machine learning? 1 point

- ○ Machine learning is the science of programming computers
- ○ Machine learning is the field of allowing robots to act intelligently.
- ○ Machine learning means from labeled data.
- ○ Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.

Curse of dimensionality: If we are able to express our inputs in few enough dimensions, we might be able to turn an unfeasible problem into a feasible one 1 point

- ○ Yes
- ○ No

Match the following

3 points

	very slow	much faster because of updates each single data point iteratively	optimise the computing infrastructure - compilers, CPUs, GPUs for performing vector additions and vector multiplications.
Batch gradient descent	○	○	○
Stochastic gradient descent (SGD)	○	○	○
mini-batch gradient descent	○	○	○

Match the following

3 points

	Statistical model	Machine learning	Artificial Intelligence
A set of techniques for estimating functions (like the one involving rent) based on datasets used for predictions of future data	○	○	○
computers behaving intelligently	○	○	○
Prediction and inference	○	○	○

Validation of a model. Choose all true statements from the following. 1 point

- [ ] K-fold cross validation more robust to over-fitting than the holdout method when performing large number of experiments.
- [ ] K-fold cross validation better to use when the dataset size is small.
- [ ] holdout method is one of the simplest cross validation methods when the dataset size is small.

Submit