EECS 545 Midterm 1 sample problems:

Essential Concepts:

1. What is the No Free Lunch theorem and why is it significant for machine learning problems?

No Free Lunch states that in order for a machine learning solution be able to generalize to unseen test data some assumptions are needed. The data alone will not be able to arrive at a solution that works well with unseen test data.

For example, it’s always possible to assign a discrete function that matches each given training point exactly, but for all unseen data points assigns a constant value C. It is possible to arbitrarily assign different values of C. Each function, or what is termed a model in machine learning, yields an error of 0 on all training data but does not allow for correct classification or prediction on unseen test data.

For generalization to take place, it is necessary to assume some other attributes, such as the training data is from a given distribution ( say Normal, or Bernoulli ),

Or that the data varies “smoothly” so extreme changes between points that are close together cannot occur. The main point here is that the data alone is not enough to allow for generalization to unseen data points )

2. For high dimensional data, what problems may occur when using linear regression with the standard normal equations ? How may this be handled ?

Two main issues may occur:

1. As the dimensionality of the data increase is becomes more likely that the normal equations, in matrix form, does not have full rank. This will make it more likely to have numeric difficulties when attempting to invert the matrix.

2. Even if the matrix form of the normal equations is not “ill conditioned”, which means it is numerically possible to compute the inverse, in high dimensions this is often computationally demanding ( computational complexity is cubic, or very nearly so with the best algorithms ) so this approach does not always scale to large dimensional data sets that well

3. Compare gradient descent with stochastic gradient descent ? What are the advantages and disadvantages of each method ?

GD: uses the full data set to determine one update to weight vector. This is accurate but often results in very slow convergence rates to obtain the solution.

SGD: perform weight update for each data point, which results in faster convergence as the weights are updated quickly. Disadvantage is that outliers or unusual data points may direct update to a local instead of global minimum, or direct the weight update away from true minimum. This is why “mini-batch” updates are used, where instead of using each data sample individually a small collection of samples ( say 10 ) is used. Note that SGD may not arrive at the same minimum point as GD and in practice it is wise to use SGD starting from random locations, in order to avoid getting stuck in a local minimum.

4. What is Bayes theorem ? Why is it commonly used ?

Bayes theorem: Posterior = ( Likelihood \* Prior ) / Evidence

Results in a full probabilistic distribution, normally used to compute the posterior distribution using the likelihood and prior. In practice, directly computation of the posterior may be difficult, while using empirical data to compute the likelihood and prior is relatively easy. In this case Bayes rule provides a practical way to determine the distribution of interest ( but not, Bayes rule may become computationally demanding in high dimensions, as to ensure this is a true distributions the “evidence” term, sometime called the “Marginal Likelihood”, may involve integrations in higher dimensions.

5. How does Logistic Regression differ from Linear Regression ? When is Logistic Regression preferred over Linear Regression ?

Linear regression is primarily used to predict a numeric value based on input data. Logistic regression is mainly used for classification, determining which class the data belongs to. Linear Regression can be used in this fashion as well, by setting threshold value, but this only results in having a class assignment and does not convey any “degree of belief of class assignment” . Simple using the predicted value from linear regression often produces a “probability of class” specified as either less than 0 or greater than one, which violates the rules of probability.

To rectify this situation, logistic regression will use an additional function, applied to the result of linear regression, to compress all possible values from the real line to the range of [0, 1] that satisfies the rules of probability.

Linear regression predicts a numeric value, whereas logistic regression is used when a probability of class membership is desired.

6. Logistic Regression often suffers from what problem when the data is linearly seperable ?

Logistic regression basically determines the probability of log ( p/ (1-p) ) where p = probability of class membership. In the linearly separable case then p can be either 0 or 1, which results in a infinite result in this calculation, which means the logistic regression algorithm will crash.

7. Classification using GDA makes what assumption about the data sets ? If the classification boundary is quadratic what does that imply about the covariance matrix ?

GDA assumes the data follows a normal distribution ( hence the G = Gaussian ). This is also an example of No Free Lunch, in order to generalize we assume the data follows a Normal distribution ( which in turn can be justified by using the Central Limit Theorem ) In the two class case, we normally assume each class has a distinct mean. If we further assume that the variance is also different for each class than the GDA classification boundary will be determined by a quadratic equation, so the boundary can follow a parabolic curve.

But, if we assume that each class has a distinct mean but the variance of each class is the same ( often done when we pool sample data together and so we can only calculate one value for variance using the available data ) then the GDA equations simplify to forcing a linear boundary condition ( so all boundaries between classes will be separated by a line )

8. What is an MLE, how does it compare to MAP ? When would you prefer to use each ?

MLE: Maximum Liklihood Estimate. The value of the likelihood function that is most probable given the data.

MAP: Maximum a Posterior: The value of the posterior estimate that is most probable given the data and an ( often assumed ) prior distribution. The distinction is MAP is a POINT estimate of the posterior distribution under the additional assumption of a prior distribution, which brings in an additional subjective element to the estimate.

Note: both MLE and MAP are point estimates, which are single values. This means we may be approximating a distribution by a single value, which makes computations faster but will contain less actual information than the full Bayesian distribution.

9. What is meant by “overfitting” ? Does having more training data increase or decrease the chances of overfitting ?

Overfitting means we create a model that fits the training data well but does not generalize to unseen test data. This means the error on training data can be made very small but the error on unseen test data does not also decrease. It results from fitting a model to the noise that is part of the data, and not the data itself. Since the noise is assumed to be random ( usually from a Normal distribution ) the more training data we use the less likely we will overfit the model.

10. How does an SVM using a hard margin differ from one using soft margins ?

Hard margin does not permit any data points to fall on the incorrect side of the margin , and therefore may not converge to a solution if a suitable boundary cannot be found. Soft margins allow some data points to fall on the wrong side of the margin. Recall, the margin is different from the boundary, a point may be on the wrong side of the margin but on the correct side of the boundary ( that is, correctly classified ). One implication, even if a hard margin solution can be found, in practice if often does not generalize as well to unseen test data ( as it is tuning itself to the training data, and in a sense this promotes overfitting, at least when compared to soft margin classifiers.

11. For SVMs, what are slack variables ?

Slack variables allow additional tuning options to permit trading off fitting an exact boundary curve that does not misclassify data vs allowing some “error” for some variables, but may produce reduced error on unseen test data. Basically, it allows for increasing the “feasible region” of any solution boundary that can be tuned, by setting the value C in the loss function, in a manner similar to regularization.

12. What is meant by Naïve Bayes ? How is it commonly used when creating an email spam filter ?

Naïve Bayes assumes that all parameters are independent, which permits calculation of the joint probability distribution using the much simpler form: the product of the individual probabilities.

When used for a spam filter, this allows calculating the probability of a word being included in a spam email or ham email, and then taking the product of the most frequently used words in known spam or ham emails and multiplying the probabilities together. This is naïve since words are often used together in messages, such as the words “Nigerian Prince” or “ Free Money” or “Great Deal!”. But in practice this naïve approach tends to work well and is relatively easy to implement.

13. True of False: Stochastic Gradient descent works better when the dimensionality of the data is high ?

False. In higher dimensions stochastic gradient descent may be prone to following the direction of the local gradient, which, since it chooses the maximum value, may not lead to the overall global function minimum. As the dimensionality increases this may happen more often (sometimes know as “ridge walking”, using the image of a function contour as a hill or ridge )

14. What is a “kernel” “

A kernel is a mathematical abstraction which can take on several meanings. In this class it is a way of calculating an inner product, which determines a measure of similarity between two vectors. Often, this permits more efficient calculations if the algorithm can be constructed using only inner products, which can then be calculated via a kernel instead of by doing the direct “dot product” in a potentially high dimensional space.

15. How do kernels make SVMs practical to implement ?

This is the “kernel trick”. If an SVM algorithm can be written so that only inner products are used, then a kernel can be substituted in for that inner product. The kernel has a computational complexity based on the original, lower dimensional feature space but computes a value that is equivalent to the value needed in the higher dimensional space. The higher dimensional space allows for a linear separation in the higher dimensional space which is not possible in lower dimensions. Kernels allow this to take place without incurring the ( potentially severe ) computational costs.

16. True of False: It is always possible to use kernels in any SVM algorithm ?

False: if the algorithm cannot be expressed solely in terms of inner products then kernels cannot be used.

17. True of False: It is always possible to use the Lagrangian dual formulation of the SVM max margin objective function to find the same value as in the primal formulation ?

False: It is always possible to construct the Langrangian dual but unless the KKT conditions hold ( which means the Gram matrix is positive semi-definite ) than the Lagrangian dual formulation is only guaranteed to produce a lower bound to the primal formulation. If the KKT conditions hold, which is an additional constraint on the problem, this means the dual formulation will definitely result in the same value for the maximum margin as would the primal formulation.

18. True or False: An SVM used to classify data will almost always require more computational time than traditional logistic regression because SVMs add extra dimensions.

False: The SVM will operate in higher dimensions but only the support values need to be calculated in order to classify new data points. Logistic Regression will use all data points in the training set and therefore may require more computation time than using SVMs.

19. What is Jensen’s Inequality ?

f(E{X}) ≤ E{f(X)}

If f(x) is a convex function, then f( Expectation of X ) is less than or equal to the Expectation of f( X ), where X is a random variable. We used this to prove that the KL divergence is greater than or equal to 0.

20. Use Jensens’s inequality to show the KL Divergence is always greater than or equal to 0 .

See notes.

21. Assume the data set is linearly separable, how does a classification algorithm using the Perceptron Learning Algorithm compare to using and SVM ?

The Perceptron algorithm will stop as soon as it finds a boundary that correctly separates the classes. An SVM will maximize the margin between the classes.

22. In Information theory, what is meant by Entropy ?

Entropy is a measure of the information contained in a message, sometimes referred to as a “measure of surprise”. Note, it is not correct to state this is a measure of randomness.

23. What is regularization ? How does L2 regularization compare to L1 regularization ?

Both are forms of coefficient shrinkage. L2 will shrink all coefficients while L1 will shrink some more than others ( and L1 often will shrink some coefficients down to 0 so it effectively performs feature selection )

24. Describe cross validation, compare k-fold cross validation with Leave One Out Cross Validation. When would you prefer using cross validation compared to simply holding out some data from your training set ( i.e. using distinct training vs test data sets )

Use of a the traditional practice of splitting available data into training data and test data requires having sufficient data available. If this is not the case, than there is a desire to use all available data for both training of the model and testing the model. But, this could easily result in overfitting. So, by dividing the data into partitions, and using each partition to both train the model a test the model, and averaging the results, we can make more efficient use of the available data and still avoid overfitting. Basically, any few data points that lead to overfitting will be averaged into the other cases and the impact of these few data points on the model is minimized, while still allowing for use of all available data.

For example, in k-fold cross validation, the available data is partitioned into k subsets, roughly equal in size, and k-1 partitions are used for training while 1 partition is used for testing. All k models are normally averaged to determine a “best fit”. In the extreme case, this partition is reduced to individual data points, which is then called Leave One Out Cross Validation ( LOOCV).

25. What is the Gram matrix ( what are it’s components )

Gram matrix K is defined as:

K = ΦΦT :

Where Knm = φ(xn)T φ(xm) = k(xn, xm) ( components are element-wise inner products of the feature vectors ) Careful here, as the Gram matrix is an outer product, where each element is calculated using inner products, and thus can be calculated by using a kernel .

Here the phi are typicaly transformed feature vectors

26. Bayes rule computation

Know Bayes rule, what is meant by posterior, likelihood, prior, and evidence, and how to use to setup a problem similar to the “Monty Hall” example.

27. Compute MLE for common distribution, Gaussian, Binomial, etc..

28. Fisher’s LDA, it tries to minimize one quantity and maximize another, what are they? How is this used as a form of dimensionality reduction?

This will find the 1-D line that best separates the data and projects onto that line.

29. The traditional optimization algorithm used with SVM is a quadratic optimization routine, but in practice often a form of gradient descent is used, why ?

Computational efficiency, particularly in higher dimensions. In practice quadratic programmers may become too slow in higher dimensions while gradient descent will not suffer the same slowdown.

30. When computing the margin for SVM, the distance from the boundary to the margin was given as + or – 1, why can this be done ?

We are solving the equation of a line: wTx +b = c. Maximizing the margin is not affected by this scaling, so we are free to fix one of the constants, in this case we can fix c= ± 1

31. Know the bias variance tradeoff

It is possible to reduce either the bias, or how fixed the form of the solution is ( for example, in simple linear regression where we fit a line to the data, we can only choose the slope and intercept, which means the solution is highly biased. Bias refers to the error in the prediction that results for choosing an incorrect form of the solution ( such as choosing a line when the true form is cubic ). Variance refers to the sensitivity of the solution to the individual variation in training data and basically tune the model to the noise in data. High bias results in underfitting while high variance results in overfitting.

32. If you are concerned about outliers affecting your classification boundary, which method among Linear Regression, Logistic Regression, or SVM would you prefer? Why?

SVM is immune to outliers since only the support vectors are used in the classification. Both Linear Regression and Logistic Regression can be sensitive to outliers.

33. Compare a generative vs a discriminant classifier, give an advantage and disadvantage of each.

A generative model allows for generating synthetic data that fits the desired distribution. A discriminative model will only yield the most likely class the data point belongs to. A generative model can provide more information about the underlying model at the cost of additional computations. Discriminant model provide less data but are less computationally demanding .

34. How does Bayesian statistics compare to Frequentist ?

A frequentist approach tries to remain objective, using only empirically obtained data. A Bayesian approach is willing to use more “subjective” information, incorporated in the form of the prior. This prior may include information such as knowledge of the underlying physical process that generates the model, or other knowledge.

One example of an approach that can use Bayesian methods that will not work with frequentist approaches is estimating the outcome of a future sporting event. A frequentist would need to see the same sporting event replayed several times, under the same conditions, in order to generate the probability estimates. This is, of course, impossible. A Bayesian could instead use “related” data, such as health of players on each team, past information regarding win/loss data, whether playing at home or on the road, etc.

In practice, machine learning makes extensive use of both approaches, though Bayesian techniques are becoming more popular.

One disadvantage of the Bayesian method is, in it’s full formulation, it requires creating a probability distribution. When using continuous variables this will result in the need to determine the value of integrals, often in high dimensions. Therefore, Bayesian methods are often very computationally demanding, and have led to approximation techniques and the of Monte Carlo methods to speed up computations.

35. What is one advantage of a full Bayesian solution to a classification or regression problem compared to using a MAP estimate ?

A MAP estimate is a point estimate, the most probable single value of the posterior given the data. It does not include as much information as does the full Bayesian distribution. The full Bayesian distribution will allow for marginalizing of variables, so if additional information is learned the Bayesian distribution can be applied to make additional inferences based on the new data.