

Quiz 3: CSCI-4202
Worth 14% of your final mark.

April 27, 2006

All of the questions require short answers (at most a few sentences). All are of equal value so answer the easy ones first!

THIS IS A OPEN BOOK, TAKE HOME QUIZZ

You may not discuss your answers with anyone!

Hand in Your Completed Quiz Under My Office Door by 9:00AM, Monday May 1!

1. What is the goal of Supervised Learning?

2. Assume that data is generated from the following function:

$$y = f(\mathbf{x}) + \rho \tag{0.1}$$

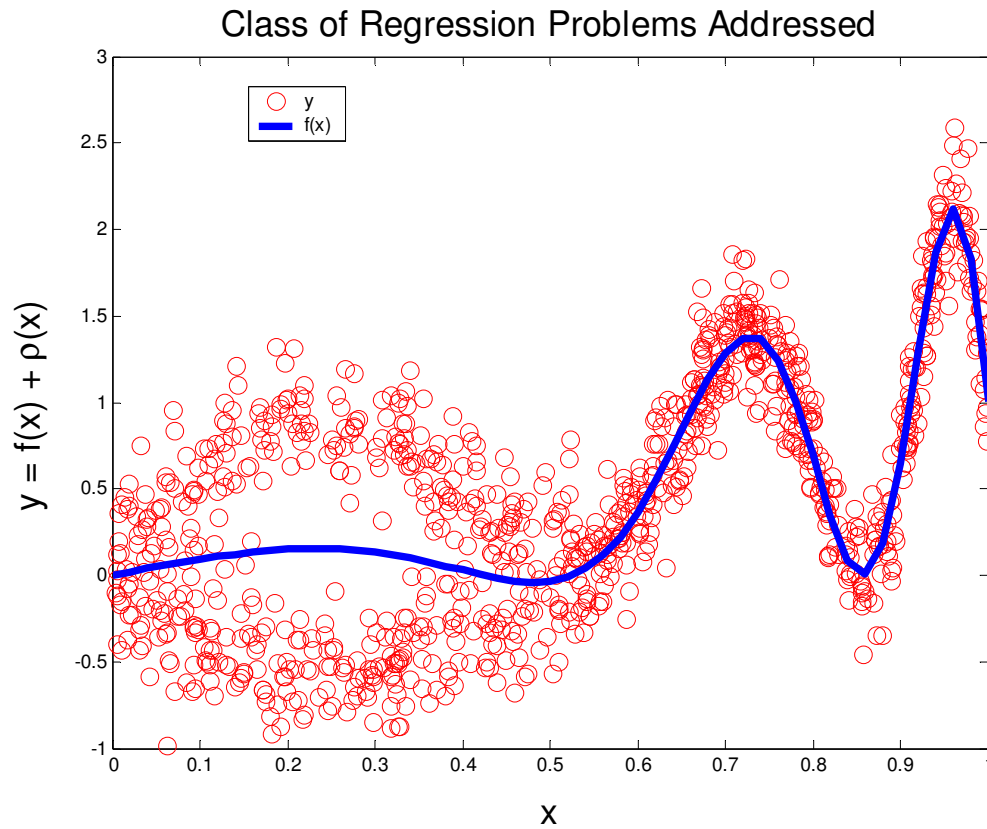
where $f(\mathbf{x}) \in \mathbb{R}$ is a real valued function, $\mathbf{x} \in \mathbb{R}^d$ is a point identically independently distributed from some stationary distribution $D(\mathbf{x})$, and ρ is a random variable with mean zero ($E[\rho] = 0$) and finite variance $V[\rho] = c$, $c > 0$, (the distribution that generates the random noise variable ρ is constant for all $\mathbf{x} \in \mathbb{R}^d$). If $\mathbf{x} = (1, 2, 0.98)$ and $c = 0.001$, how many different values of y can be observed (note that the notation $\mathbf{x} = (1, 2, 0.98)$ means that \mathbf{x} is a three dimensional vector and hence $f(\mathbf{x})$ is a function of three variables)?

How many different values of $f(\mathbf{x})$ can be observed when $\mathbf{x} = (1, 2, 0.98)$ and $c = 0.001$?

Similarly, if $\mathbf{x} = (0.29, 0.11, 2.1)$ and $c = 0$, how many different values of y can be observed?

How many different values of $f(\mathbf{x})$ can be observed when $\mathbf{x} = (0.29, 0.11, 2.1)$ and $c = 0$?

3. Is the data given in the figure below consistent with the assumptions in the previous question – briefly explain your answer?



4. Assume you would like to learn a linear model of the form

$$y = \sum_{i=1}^d \beta_i x_i + \beta_0$$

Assume that the coefficients β_0, \dots, β_N are obtained from the learning data $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)$ to minimize:

$$\hat{\beta} = \arg \min_{\beta} \left\{ \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^d \beta_j x_{ij} \right)^2 \right\}$$

What is this type of learning algorithm called?

5. Assume the same linear model as in the previous question. However, now the coefficients β_0, \dots, β_N are obtained from the learning data $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)$ to minimize:

$$\hat{\beta} = \arg \min_{\beta} \left\{ \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^d \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^d \beta_j^2 \right\}$$

where $\lambda \geq 0$. What is this type of learning algorithm called?

What happens when you set $\lambda = 0$?

What happens to the coefficients β_1, \dots, β_N as λ is increased?

When would you use this learning algorithm over the one in the previous question?

6. Assume the same linear model as in the previous two questions. However, now the coefficients β_0, \dots, β_N are obtained from the learning data $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)$ to minimize:

$$\hat{\beta} = \arg \min_{\beta} \left\{ \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^d \beta_j x_{ij} \right)^2 \right\}$$

subject to: $\sum_{j=1}^d |\beta_j| \leq s, \quad s > 0$

What is this type of learning algorithm called?

When would you use this learning algorithm over the one in the previous question?

7. Now assume a model with the following structure:

$$y = \sum_{i=1}^K \beta_i \phi_i(\mathbf{x}) + \beta_0$$

where $\phi_i(\mathbf{x})$ are nonlinear basis functions of \mathbf{x} (e.g. $\phi_i(\mathbf{x}) = x_1 x_2$ is one example). Is this a linear model in basis function space $\phi_1(\mathbf{x}), \dots, \phi_K(\mathbf{x})$?

Is this a linear model in $\mathbf{x} \in \mathbb{R}^d$ space?

8. You would like to learn a linear model of the form

$$y = \sum_{i=1}^d \beta_i x_i + \beta_0$$

Assume that the coefficients β_0, \dots, β_N are obtained from the learning data $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)$ using ridge regression, and that $\mathbf{x} \in \mathbb{R}^d$, $d = 1000000$ and $N = 10$. Would it be more computationally efficient to use Kernel Ridge Regression or standard Linear Ridge Regression to estimate β_0, \dots, β_N ? Why?

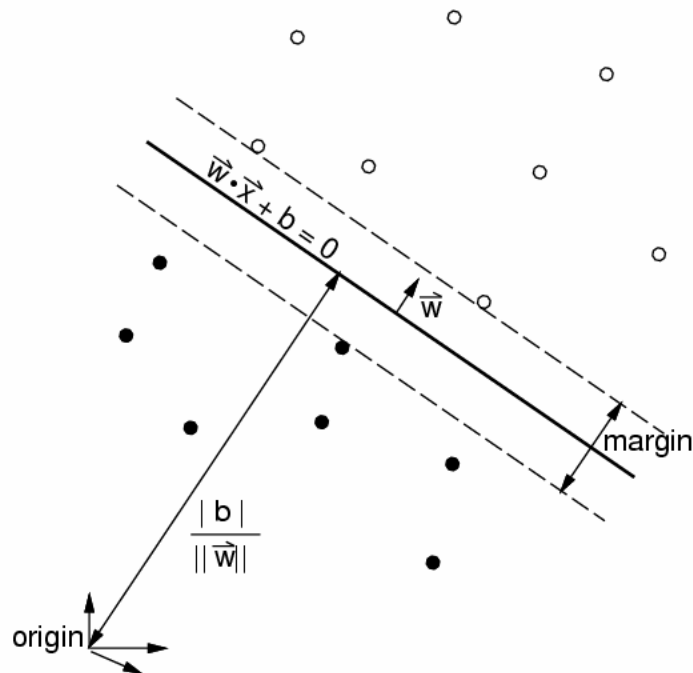
9. If the model is $\hat{y} = 2x_1^2 + 3x_2^9 - 5$, what is \hat{y} when $x_1 = 1$ and $x_2 = 1$?

10. Give an example (i.e. write out a description of the algorithm) of the K Nearest Neighbor algorithm when $K=1$.

11. Assume a loss function $L(a_1, \dots, a_K)$ that depends on K parameters a_1, \dots, a_K , and assume that it is differentiable with respect to those parameters. Give an update formula for incrementally modifying the a_1, \dots, a_K such that the loss function value will decrease?

12. Under what conditions will the algorithm you defined in the previous question converge to a globally optimal solution?

13. Circle the Support Vectors in the following 2-D data:

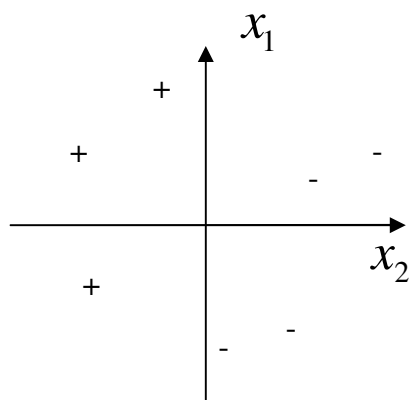


What quantity is maximized to obtain these support vectors?

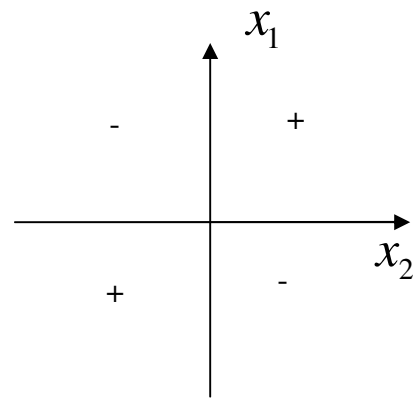
What is the effect of maximizing this quantity?

Do support vector machines find a global maximum for this quantity?

14. Below are plots of two data sets (each having two classes: + and -). Which data set is linearly separable (i.e. circle the linearly separable set)?



a)



b)

15. For Nonlinear Support Vector Machines, input data is projected into what nonlinear space?

Is it the same space for Classification and Regression Support Vector Machines?

16. Assume that I have the following support vector classification model

$$\hat{y} = f(\mathbf{x}) = \text{sgn}[5 - 4.2 \cdot K((1, 3), \mathbf{x}) + 2.1 \cdot K((-3, 2), \mathbf{x})]$$

(note that $(1, 3)$ and $(-3, 2)$ are 2 dimensional vectors) From this model, can you identify the number of training examples used to construct the model?

Can you give the (\mathbf{x}_i, y_i) values for any of the training examples used to construct this model?

17. Define a kernel matrix and give pseudo-code for evaluating it.

18. What is the difference between classification and regression data?

19. Can I use a regression algorithm to solve a classification problem? If your answer is yes, describe how it can be done? If it is no, describe why not?

20. Can I use a classification algorithm to solve a regression problem? If your answer is yes, describe how it can be done? If it is no, describe why not?

21. Define N Fold cross validation.

22. If I use N Fold cross validation to select learning parameters, is the error rate that is returned by the N Fold cross validation procedure a good indication of how well the model built by my algorithm will do on future data? If not, what is? (in other words, how can I compare how well two different learning algorithms will do on a specific data set?)
23. What is an unstable learning algorithm? Give two examples of unstable predictors. Give one example of a stable predictor.
24. How is Bagging different from Deterministic Boosting?

25. If you are using single tree stumps as the base classifier in Deterministic Boosting, could you build a classifier that separates the following data? (Assume that the stumps are constructed one at a time using a single variable, and that the split variable chosen is based on minimizing the entropy after the split)

x_1	x_2	x_2	y
0	0	0	-1
0	0	1	+1
0	1	0	+1
0	1	1	-1
1	0	0	+1
1	0	1	-1
1	1	0	-1
1	1	1	+1

26. How does bagging of classification trees compare to the random forests algorithm?