1. **True or False** Please give an explanation for your answer, this is worth 1 pt/question.
   (a) (2 points) No classifier can do better than a naive Bayes classifier if the distribution of the data is known.

   (b) (2 points) Maximizing the likelihood of linear regression yields multiple local optimums.

   (c) (2 points) If a function is not twice differentiable, that is the Hessian is undefined, then it cannot be convex.

   (d) (2 points) Ridge regression; linear regression with the l_2 penalty is a convex function.

2. **Short Answer**
   (a) (2 points) Explain how you would use 10-fold cross validation to choose λ for l_1-regularized linear regression.

   (b) (2 points) Why does the kernel trick allow us to solve SVMs with high dimensional feature spaces, without significantly increasing the running time.

   (c) (5 points) Consider the optimization problem

$$\begin{aligned}&\text{minimize } x^2 + 1\\&\text{subject to } (x-2)(x-4) \leq 0\end{aligned}$$

   State the dual problem.

9 Naive Bayes [10 pts]

Given the following training (x, y), what problem will Naive Bayes encounter with test data $z2$?

```
x1 = (0,0,0,1,0,0,1)     y1 =  1
x2 = (0,0,1,1,0,0,0)     y2 =  1
x3 = (1,1,0,0,0,1,0)     y3 = -1
x4 = (1,0,0,0,1,1,0)     y4 = -1

z1 = (1,0,0,0,0,1,0)
z2 = (0,1,1,0,0,1,1)
```

Using your fix for the problem, compute the Naive Bayes estimate for $z1$ and $z2$.

10 Perceptron [10 pts]

Demonstrate how the perceptron without bias (i.e. we set the parameter $b = 0$ and keep it fixed) updates its parameters given the following training sequence:

```
x1 = (0,0,0,1,0,0,1)     y1 =  1
x2 = (1,1,0,0,0,1,0)     y2 = -1
x3 = (0,0,1,1,0,0,0)     y3 =  1
x4 = (1,0,0,0,1,1,0)     y4 = -1
x5 = (1,0,0,0,0,1,0)     y5 = -1
```

2 [16 Points] SVMs and the slack penalty C

The goal of this problem is to correctly classify test data points, given a training data set. You have been warned, however, that the training data comes from sensors which can be error-prone, so you should avoid trusting any specific point too much.

For this problem, assume that we are training an SVM with a **quadratic kernel**– that is, our kernel function is a polynomial kernel of degree 2. You are given the data set presented in Figure 1. The slack penalty C will determine the location of the separating hyperplane. Please answer the following questions *qualitatively*. Give a one sentence answer/justification for each and draw your solution in the appropriate part of the Figure at the end of the problem.

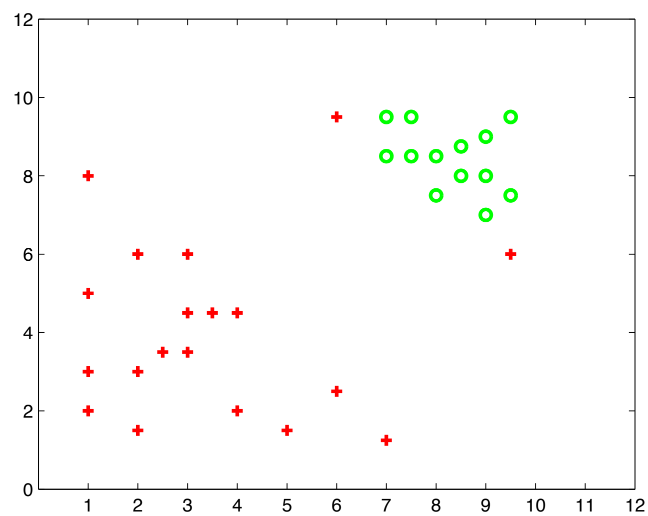


Figure 1: Dataset for SVM slack penalty selection task in Question 2.

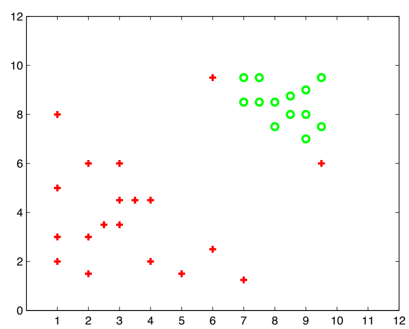
1. [4 points] Where would the decision boundary be for very large values of C (i.e., $C \to \infty$)? (remember that we are using an SVM with a quadratic kernel.) Draw on the figure below. Justify your answer.

2. [4 points] For $C \approx 0$, indicate in the figure below, where you would expect the decision boundary to be? Justify your answer.

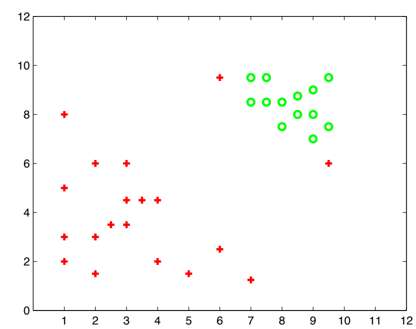
3. [2 points] Which of the two cases above would you expect to work better in the classification task? Why?

4. [3 points] Draw a data point which will not change the decision boundary learned for very large values of C. Justify your answer.

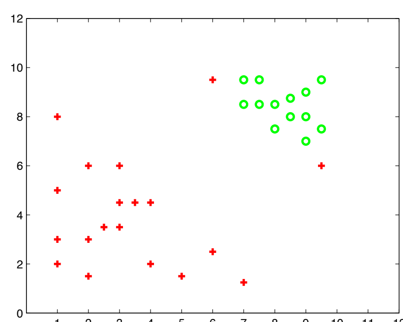
5. [3 points] Draw a data point which will significantly change the decision boundary learned for very large values of C. Justify your answer.



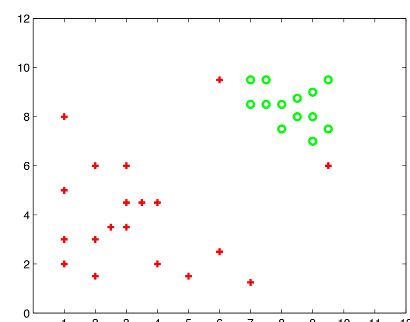
(a) Part 1



(b) Part 2



(c) Part 4



(d) Part 5

Figure 2: Draw your solutions for Problem 2 here.

1 Conditional Independence, MLE/MAP, Probability (12 pts)

1. (4 pts) Show that $\Pr(X, Y|Z) = \Pr(X|Z)\Pr(Y|Z)$ if $\Pr(X|Y, Z) = \Pr(X|Z)$.

2. (4 pts) If a data point y follows the Poisson distribution with rate parameter θ, then the probability of a single observation y is

$$p(y|\theta) = \frac{\theta^y e^{-\theta}}{y!}, \quad \text{for } y = 0, 1, 2, \cdots.$$

   You are given data points $y_1, \cdots, y_n$ independently drawn from a Poisson distribution with parameter θ. Write down the log-likelihood of the data as a function of θ.

3. (4 pts) Suppose that in answering a question in a multiple choice test, an examinee either knows the answer, with probability p, or he guesses with probability $1 - p$. Assume that the probability of answering a question correctly is 1 for an examinee who knows the answer and $1/m$ for the examinee who guesses, where m is the number of multiple choice alternatives. What is the probability that an examinee knew the answer to a question, given that he has correctly answered it?

4 Bias-Variance Decomposition (12 pts)

1. (6 pts) Suppose you have regression data generated by a polynomial of degree 3. Characterize the bias-variance of the estimates of the following models on the data with respect to the true model by circling the appropriate entry.

	Bias	Variance
Linear regression	low/high	low/high
Polynomial regression with degree 3	low/high	low/high
Polynomial regression with degree 10	low/high	low/high

2. Let $Y = f(X) + \epsilon$, where ϵ has mean zero and variance σ_ϵ^2. In k-nearest neighbor (kNN) regression, the prediction of Y at point x_0 is given by the average of the values Y at the k neighbors closest to x_0.

   (a) (2 pts) Denote the ℓ-nearest neighbor to x_0 by $x_{(\ell)}$ and its corresponding Y value by $y_{(\ell)}$. Write the prediction $\hat{f}(x_0)$ of the kNN regression for x_0 in terms of $y_{(\ell)}, 1 \leq \ell \leq k$.

   (b) (2 pts) What is the behavior of the bias as k increases?

   (c) (2 pts) What is the behavior of the variance as k increases?

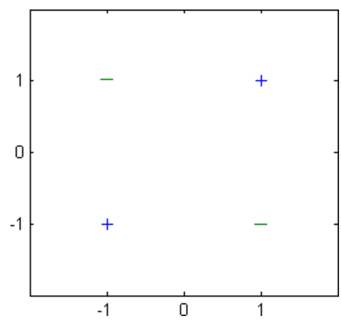
5 Support Vector Machine (12 pts)

Consider a supervised learning problem in which the training examples are points in 2-dimensional space. The positive examples are $(1,1)$ and $(-1,-1)$. The negative examples are $(1,-1)$ and $(-1,1)$.

1. (1 pts) Are the positive examples linearly separable from the negative examples in the original space?

2. (4 pts) Consider the feature transformation $\phi(x) = [1, x_1, x_2, x_1 x_2]$, where x_1 and x_2 are, respectively, the first and second coordinates of a generic example x. The prediction function is $y(x) = w^T * \phi(x)$ in this feature space. Give the coefficients, w, of a maximum-margin decision surface separating the positive examples from the negative examples. (You should be able to do this by inspection, without any significant computation.)

3. (3 pts) Add one training example to the graph so the total five examples can no longer be linearly separated in the feature space $\phi(x)$ defined in problem 5.2.



4. (4 pts) What kernel $K(x, x')$ does this feature transformation ϕ correspond to?

6 Generative vs. Discriminative Classifier (20 pts)

Consider the binary classification problem where class label $Y \in \{0,1\}$ and each training example X has 2 binary attributes $X_1, X_2 \in \{0,1\}$.

In this problem, we will always assume X_1 and X_2 are conditional independent given Y, that the class priors are $P(Y=0) = P(Y=1) = 0.5$, and that the conditional probabilities are as follows:

$P(X_1 \mid Y)$	$X_1 = 0$	$X_1 = 1$
$Y=0$	0.7	0.3
$Y=1$	0.2	0.8

$P(X_2 \mid Y)$	$X_2 = 0$	$X_2 = 1$
$Y=0$	0.9	0.1
$Y=1$	0.5	0.5

The expected error rate is the probability that a classifier provides an incorrect prediction for an observation: if Y is the true label, let $\hat{Y}(X_1, X_2)$ be the predicted class label, then the expected error rate is

$$P_{\mathcal{D}}\left(Y = 1 - \hat{Y}(X_1, X_2)\right) = \sum_{X_1=0}^{1} \sum_{X_2=0}^{1} P_{\mathcal{D}}\left(X_1, X_2, Y = 1 - \hat{Y}(X_1, X_2)\right).$$

Note that we use the subscript $\mathcal{D}$ to emphasize that the probabilities are computed under the true distribution of the data.

*You don't need to show all the derivation for your answers in this problem.

1. (4 pts) Write down the naïve Bayes prediction for all the 4 possible configurations of X_1, X_2. The following table would help you to complete this problem.

X_1	X_2	$P(X_1, X_2, Y=0)$	$P(X_1, X_2, Y=1)$	$\hat{Y}(X_1, X_2)$
0	0			
0	1			
1	0			
1	1			

2. (4 pts) Compute the expected error rate of this naïve Bayes classifier which predicts Y given both of the attributes $\{X_1, X_2\}$. Assume that the classifier is learned with infinite training data.

3 Logistic Regression [18 pts]

We consider here a discriminative approach for solving the classification problem illustrated in Figure 1.

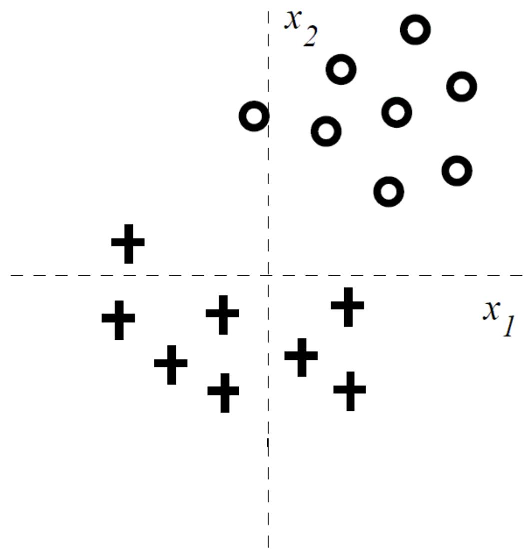


Figure 1: The 2-dimensional labeled training set, where '+' corresponds to class y=1 and 'O' corresponds to class $y = 0$.

1. We attempt to solve the binary classification task depicted in Figure 1 with the simple linear logistic regression model

$$P(y = 1|\vec{x}, \vec{w}) = g(w_0 + w_1x_1 + w_2x_2) = \frac{1}{1 + exp(-w_0 - w_1x_1 - w_2x_2)}.$$

   Notice that the training data can be separated with *zero* training error with a linear separator.

   Consider training *regularized* linear logistic regression models where we try to maximize

$$\sum_{i=1}^{n} \log\ (P(y_i|x_i, w_0, w_1, w_2)) - Cw_j^2$$

   for very large C. The regularization penalties used in penalized conditional log-likelihood estimation are $-Cw_j^2$, where $j = \{0, 1, 2\}$. In other words, only one of the parameters is regularized in each case. Given the training data in Figure 1, how does the training error change with regularization of each parameter w_j? State whether the training error increases or stays the same (zero) for each w_j for very large C. Provide a brief justification for each of your answers.

(a) By regularizing w_2 **[3 pts]**

(b) By regularizing w_1 **[3 pts]**

(c) By regularizing w_0 **[3 pts]**

2. If we change the form of regularization to L1-norm (absolute value) and regularize w_1 and w_2 only (but not w_0), we get the following penalized log-likelihood

$$\sum_{i=1}^{n} \log P(y_i|x_i, w_0, w_1, w_2) - C(|w_1| + |w_2|).$$

Consider again the problem in Figure 1 and the same linear logistic regression model $P(y = 1|\vec{x}, \vec{w}) = g(w_0 + w_1x_1 + w_2x_2)$.

(a) **[3 pts]** As we increase the regularization parameter C which of the following scenarios do you expect to observe? (Choose only one) Briefly explain your choice:

( ) First w_1 will become 0, then w_2.

( ) First w_2 will become 0, then w_1.

( ) w_1 and w_2 will become zero simultaneously.

( ) None of the weights will become exactly zero, only smaller as C increases.

(b) **[3 pts]** For very large C, with the same L1-norm regularization for w_1 and w_2 as above, which value(s) do you expect w_0 to take? Explain briefly. (Note that the number of points from each class is the same.) (You can give a range of values for w_0 if you deem necessary).

(c) **[3 pts]** Assume that we obtain more data points from the '+' class that corresponds to y=1 so that the class labels become unbalanced. Again for very large C, with the same L1-norm regularization for w_1 and w_2 as above, which value(s) do you expect w_0 to take? Explain briefly. (You can give a range of values for w_0 if you deem necessary).

4 Kernel regression [16 pts]

Now lets consider the non-parametric kernel regression setting. In this problem, you will investigate univariate locally linear regression where the estimator is of the form:

$$\widehat{f}(x) = \beta_1 + \beta_2 x$$

and the solution for parameter vector $\beta = [\beta_1 \ \beta_2]$ is obtained by minimizing the weighted least square error:

$$J(\beta_1, \beta_2) = \sum_{i=1}^{n} W_i(x)(Y_i - \beta_1 - \beta_2 X_i)^2 \quad \text{where} \quad W_i(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)},$$

where K is a kernel with bandwidth h. Observe that the weighted least squares error can be expressed in matrix form as

$$J(\beta_1, \beta_2) = (Y - A\beta)^T W (Y - A\beta),$$

where Y is a vector of n labels in the training example, W is a $n \times n$ diagonal matrix with weight of each training example on the diagonal, and

$$A = \begin{bmatrix} 1 \ X_1 \\ 1 \ X_2 \\ \dots \\ 1 \ X_n \end{bmatrix}$$

1. [**4 pts**] Derive an expression in matrix form for the solution vector $\hat{\beta}$ that minimizes the weighted least square.

2. [**3 pts**] When is the above solution unique?

3. [**3 pts**] If the solution is not unique, one approach is to optimize the objective function J using gradient descent. Write the update equation for gradient descent in this case. Note: Your answer must be expressed in terms of the matrices defined above.

4. [**3 pts**] Can you identify the signal plus noise model under which maximizing the likelihood (MLE) corresponds to the weighted least squares formulation mentioned above?

5. [**3 pts**] Why is the above setting non-parametric? Mention one advantage and one disadvantage of nonparametric techniques over parametric techniques.