6.867 Machine learning

Mid-term exam

October 17, 2007

(2 points) Your name and MIT ID:

Problem 1

Figure 1 plots SVM decision boundaries resulting from using different kernels and/or different slack penalties. The methods used to generate the plots are listed below but (the absent minded) professor forgot to label them. Please assign the plots to the right method. Oh, we also forgot to list one of the methods.

1.1 **(2 points)** $\min \frac{1}{2}\|\underline{\theta}\|^2 + C\sum_{t=1}^{n}\xi_t$ s.t.

$$\xi_t \geq 0, \quad y_t(\underline{\theta}^T\underline{x}_t + \theta_0) - 1 + \xi_t \geq 0, \quad t = 1, \ldots, n$$

where $C = 0.1$.

1.2 **(2 points)** $\min \frac{1}{2}\|\underline{\theta}\|^2 + C\sum_{t=1}^{n}\xi_t$ s.t.

$$\xi_t \geq 0, \quad y_t(\underline{\theta}^T\underline{x}_t + \theta_0) - 1 + \xi_t \geq 0, \quad t = 1, \ldots, n$$

where $C = 1$.

1.3 **(2 points)**

$$\max \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j K(\underline{x}_i, \underline{x}_j)$$

$$\alpha_i \geq 0, \; i = 1, \ldots, n, \quad \sum_{i=1}^{n}\alpha_i y_i = 0$$

where $K(\underline{x}, \underline{x}') = \underline{x}^T\underline{x}' + (\underline{x}^T\underline{x}')^2$.

1.4 **(2 points)**

$$\max \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j K(\underline{x}_i, \underline{x}_j)$$

$$\alpha_i \geq 0, \; i = 1, \ldots, n, \quad \sum_{i=1}^{n}\alpha_i y_i = 0$$

where $K(\underline{x}, \underline{x}') = \exp(-1/2\|\underline{x} - \underline{x}'\|^2)$.

1.5 **(2 points)**

$$\max \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j K(\underline{x}_i, \underline{x}_j)$$

$$\alpha_i \geq 0, \; i = 1, \ldots, n, \quad \sum_{i=1}^{n}\alpha_i y_i = 0$$

where $K(\underline{x}, \underline{x}') = \exp(-\|\underline{x} - \underline{x}'\|^2)$ (only the kernel is different from 1.4)

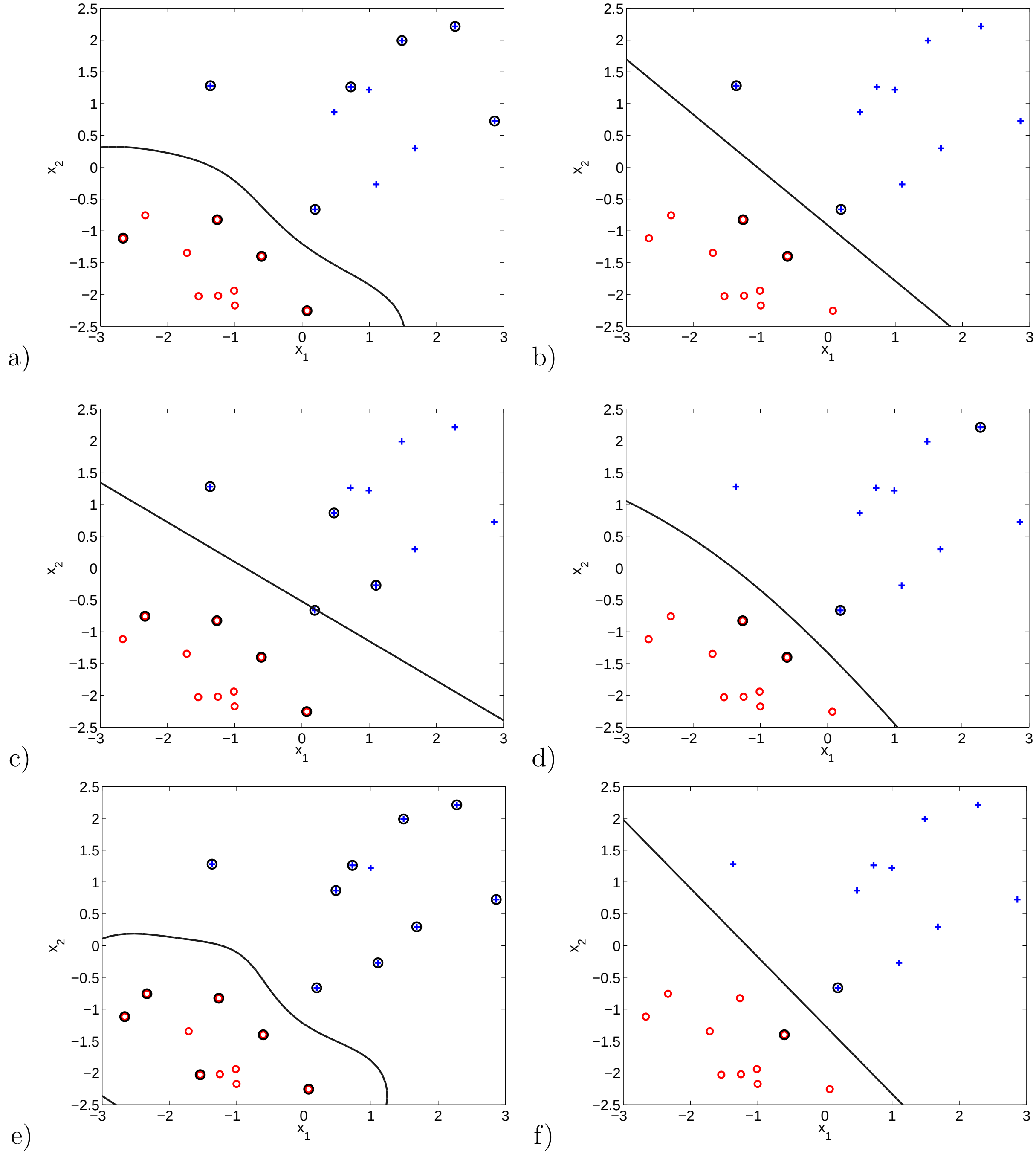

Figure 1: plots of SVM decision boundaries with different kernels and/or slack penalties

1.6 **(4 points)** Consider the linear SVM with slack penalties

$$\min \ \frac{1}{2}\|\underline{\theta}\|^2 + C\sum_{t=1}^{n}\xi_t \ \text{ s.t.}$$

$$\xi_t \geq 0, \ \ y_t(\underline{\theta}^T\underline{x}_t + \theta_0) - 1 + \xi_t \geq 0, \ \ t = 1, \ldots, n$$

Indicate which of the following statements hold as we *increase* the parameter C from any starting value. Use 'Y' for statements that *will necessarily hold*, 'N' if the statement is *never true*, and 'D' if the validity of the statement depends on the situation when C increases.

( ) θ_0 will not increase

( ) $\|\underline{\hat{\theta}}\|$ increases

( ) $\|\underline{\hat{\theta}}\|$ will not decrease

( ) more points will be misclassified

( ) the geometric margin will not increase

Problem 2

We are interested in modeling the relationship between real inputs x and responses y. We will use a simple linear regression model for this purpose. So, according to our model

$$y = \theta_1 x + \theta_0 + \epsilon = \underline{\beta}^T \begin{bmatrix} x \\ 1 \end{bmatrix} + \epsilon$$

where $\underline{\beta} = [\theta_1, \theta_0]^T$ and $\epsilon \sim N(0, \sigma^2)$. We were a bit unlucky in choosing our model, however, since the inputs and responses are actually related quadratically:

$$y = \theta_2^* x^2 + \theta_1^* x + \theta_0^* + \epsilon = \underline{\beta}^{*T} \begin{bmatrix} x^2 \\ x \\ 1 \end{bmatrix} + \epsilon$$

where $\underline{\beta}^* = [\theta_2^*, \theta_1^*, \theta_0^*]^T$ and $\epsilon \sim N(0, \sigma^{*2})$. In other words, we are modeling the underlying and *unknown* quadratic relation with a linear model.

In a context of a specific training set of inputs $x_1, \ldots, x_n$ and responses $y_1, \ldots, y_n$, we define

$$X_n = \begin{bmatrix} x_1 & 1 \\ \cdots & \cdots \\ x_n & 1 \end{bmatrix}, \quad X_n^* = \begin{bmatrix} x_1^2 & x_1 & 1 \\ \cdots & \cdots & \cdots \\ x_n^2 & x_n & 1 \end{bmatrix}, \quad \underline{y} = \begin{bmatrix} y_1 \\ \cdots \\ y_n \end{bmatrix}$$

The least squares estimates for the parameters in our model are then given by

$$\underline{\hat{\beta}} = (X_n^T X_n)^{-1} X_n^T \underline{y}, \quad \underline{\hat{\beta}}^T = \underline{y}^T X_n (X_n^T X_n)^{-1}$$

2.1 **(2 points)** What is the predicted response $\hat{y}(x)$ from our model at a new point x?

2.2 **(3 points)** Write down an expression for the bias of $\hat{y}(x)$ at a fixed input x when the expectation is taken over the possible responses $y_1, \ldots, y_n$ for fixed training inputs $x_1, \ldots, x_n$. (the final expression should not involve expectations)

2.3 **(4 points)** Specify a possible training set with five points in Figure 2.3 that illustrates why the predicted responses cannot be expected to be unbiased for all x in our setting. Indicate a rough value of σ^* that you are assuming for your sampled training data.

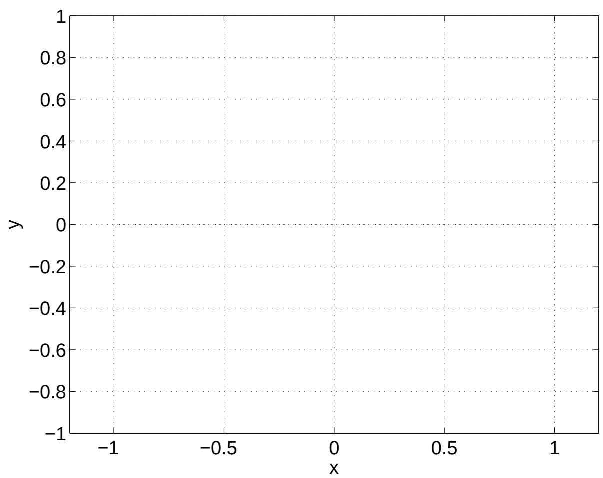

Figure 2.3: Your answer training set for problem 2.3.

2.4 **(3 points)** Which of the following input selection criteria are likely to work in our setting in terms of leading to the best linear approximation? Assume that $x \in [-1, 1]$.

a) ( ) Randomly select each x from within the interval $[-1, 1]$

b) ( ) Sequentially select points so as to minimize the trace of $(X_n^T X_n)^{-1}$

c) ( ) Select the next input to be x that maximizes the mean squared prediction error

$$E\{(\hat{y}(x) - y^*(x))^2 \,|x\}$$

(3 points) Briefly justify your answer to part c)

Problem 3

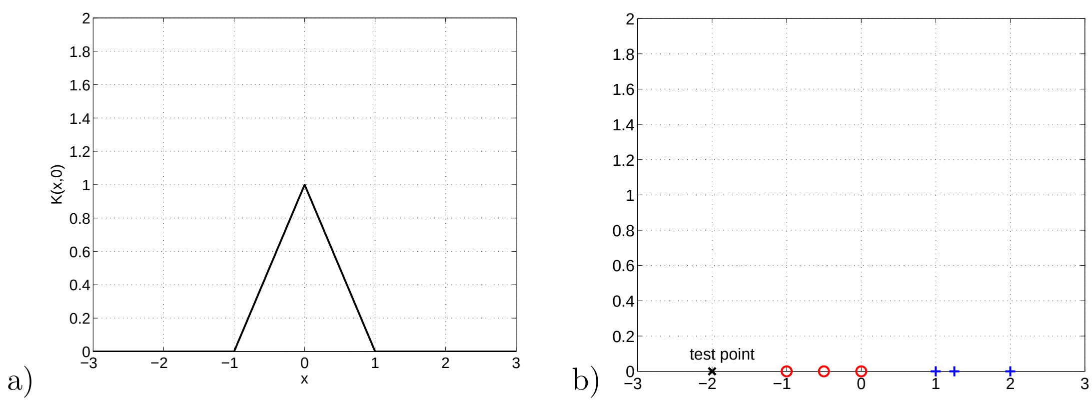

Figure 3: a) Kernel $K(x, 0)$ for problem 3. b) data for problems 3.2 and 3.3.

Consider solving a 1-dimensional classification problem with SVMs and the kernel

$$K(x, x') = (1 - |x - x'|)^+ = \max\{0,\ 1 - |x - x'|\}$$

Figure 3a) illustrates this kernel $K(x, 0)$ as a function of x. The feature "vectors" corresponding to this kernel are actually functions $\phi(\cdot\,; x)$ such that

$$K(x, x') = \int_{-\infty}^{\infty} \phi(z\,; x)\phi(z\,; x')dz$$

3.1 **(2 points)** What is the value of $\|\phi(\cdot\,; x)\|$ at $x = 0$?

3.2 **(3 points)** What is the dual objective function for training SVMs (no slack) when we do not include the offset term θ_0 in the classifier? We maximize

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

subject to ?

3.3 **(3 points)** What is the value of the discriminant function $\sum_{i \in n} \hat{\alpha}_i y_i K(x, x_i)$ on the test point in Figure 3b)? Assume that $\hat{\alpha}_i$ are estimated on the basis of the training data in the figure without an offset parameter.

3.4 **(2 points)** Would the test point in Figure 3b) become a support vector if it were included in the training set?

3.5 **(2 points)** We can improve the kernel function a bit by introducing a width parameter σ such that

$$K(x, x') = (1 - |x - x'|/\sigma)^{+}$$

What would be a reasonable method for choosing σ?

3.6 **(4 points)** Would your method solve the problem identified in 3.3? Briefly explain why or why not.

3.7 **(4 points)** It is sometimes useful to incorporate test inputs (if available) in some manner in training the classifier. How could you include the test points in selecting the kernel width parameter σ?

Problem 4

We consider here a logistic regression model for classifying midterm exams. The class labels indicate whether the exam is good ($y = 1$) or bad ($y = -1$). The probabilities over the labels, given the exam x, are assigned according to

$$P(y = 1|x, \underline{\theta}) = g(\,\underline{\theta}^T \underline{\phi}(x)\,)$$

where $g(z) = (1 + e^{-z})^{-1}$ is the logistic function. The feature vectors simply indicate whether a word w appears in the exam x:

$$\phi_w(x) = \begin{cases} 1, & \text{if } x \text{ contains word } w \\ 0, & \text{otherwise} \end{cases}$$

There are only two words we are interested in so that $w \in \{\text{svm}, \text{kernel}\}$. The exams are first turned into all lowercase letters before evaluating the corresponding feature vectors.

We would like to train the logistic regression model based on past exams $x_1, \ldots, x_n$ and labels $y_1, \ldots, y_n$ (from student ratings) by maximizing the penalized log-likelihood of the labels:

$$\sum_{t=1}^{n} \log P(y_t|x_t, \underline{\theta}) - \frac{\lambda}{2}\|\underline{\theta}\|^2 = \sum_{t=1}^{n} \log g(\, y_t\underline{\theta}^T \underline{\phi}(x_t)\,) - \frac{\lambda}{2}\|\underline{\theta}\|^2$$

The problem is a bit hard to solve well, however, since we only have three labeled exams:

$$\begin{array}{ll} \underline{\phi}(x_1) = [1,1]^T & y_1 = 1 \\ \underline{\phi}(x_2) = [1,0]^T & y_2 = -1 \\ \underline{\phi}(x_3) = [0,0]^T & y_3 = 1 \end{array}$$

4.1 **(2 points)** Does it matter how the third exam is labeled? (Y/N)

4.2 **(2 points)** What would be the value of the resulting training log-likelihood be if we set $\lambda = 0$?

4.3 **(2 points)** The logistic regression model associates class probabilities with each point. Does the effect of the regularization penalty on these probabilities depend on the norms $\|\underline{\phi}(x_t)\|$? (Y/N)

4.4 **(4 points)** For large λ (strong regularization), the log-likelihood terms will behave as linear functions of $\underline{\theta}$ (see Figure 4).

$$\log g(\, y_t\underline{\theta}^T \underline{\phi}(x_t)\,) \approx \frac{1}{2} y_t\underline{\theta}^T \underline{\phi}(x_t)$$

In this regime (large λ), draw in Figure 4 how $\hat{\underline{\theta}}$ behaves as a function of λ. In other words, draw $\hat{\underline{\theta}}$ (at any scale) and its direction of change with increasing λ. We will classify correctly only one of the training examples. Why?

4.5 **(3 points)** For general $\lambda > 0$, will the resulting classification decisions (predicted labels) for new exams depend on the value of λ? (Y/N)

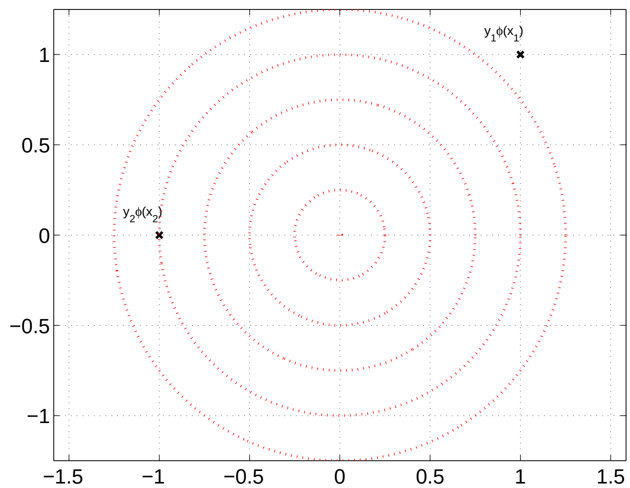

Figure 4: Points $y_1\underline{\phi}(x_1)$ and $y_2\underline{\phi}(x_2)$ along with the contours of the regularization term $\|\underline{\theta}\|$.

Additional set of figures

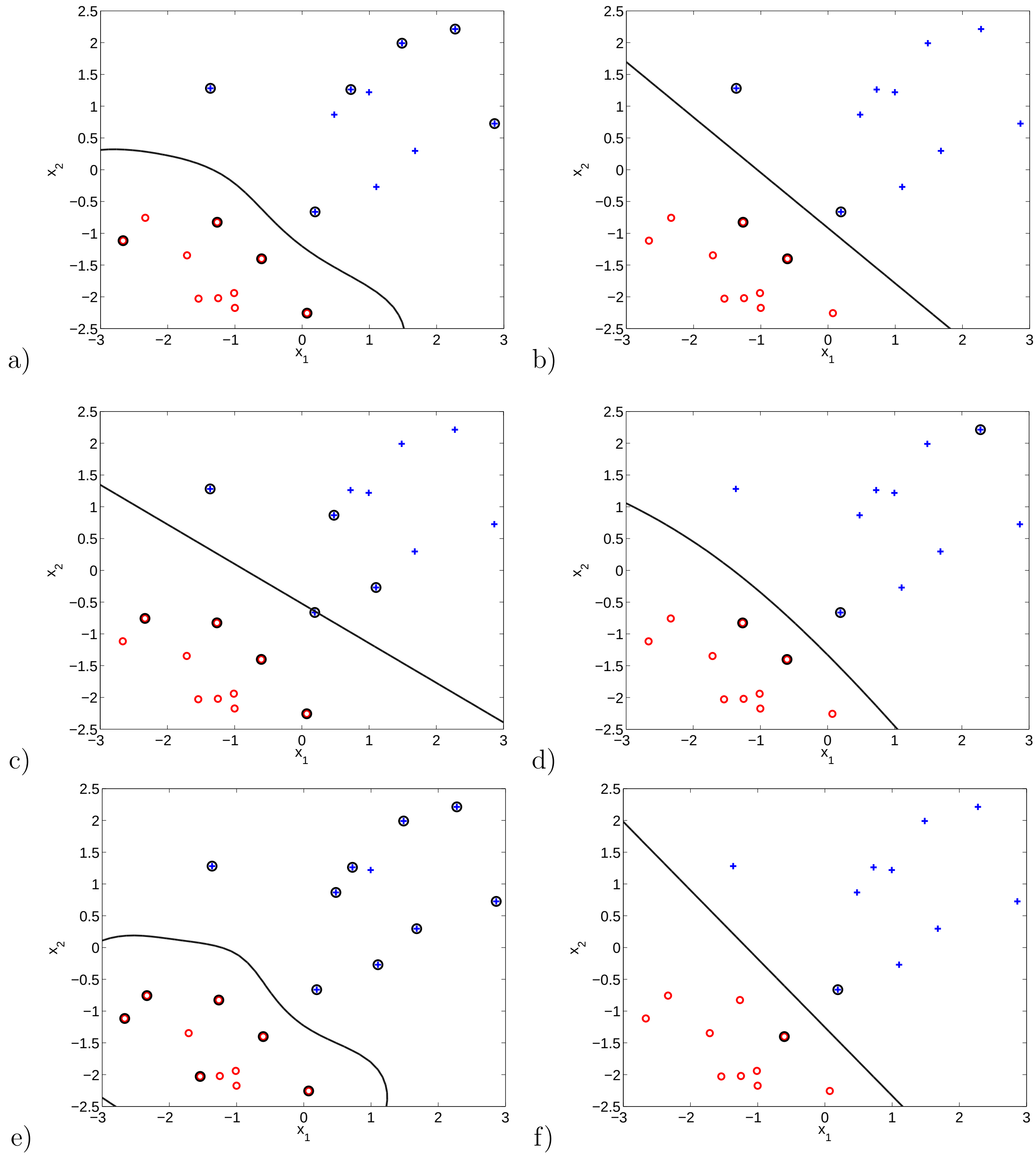

Figure 1: plots of SVM decision boundaries with different kernels and/or slack penalties

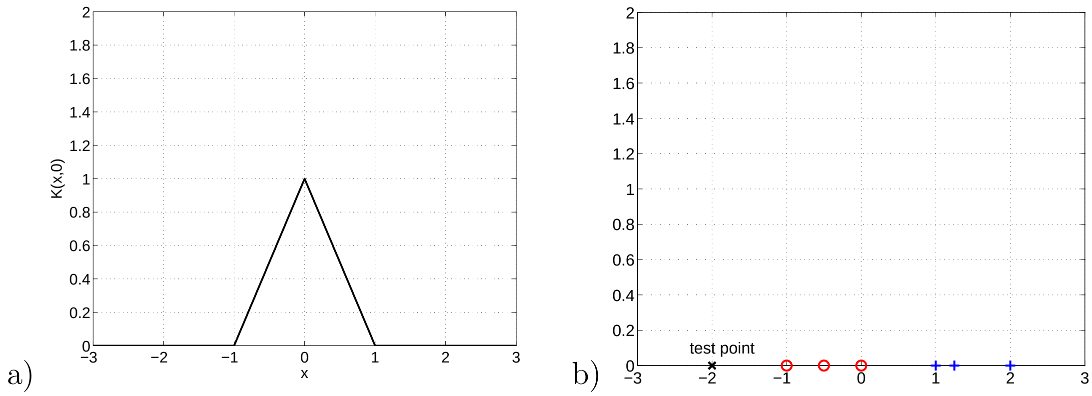

Figure 3: a) Kernel $K(x,0)$ for problem 3. b) data for problems 3.2 and 3.3.

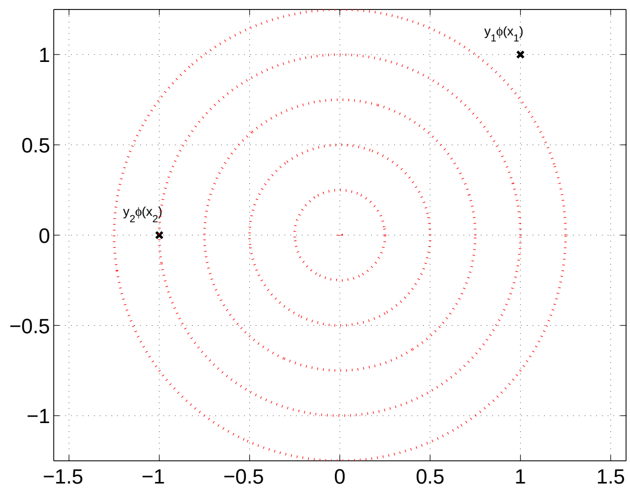

Figure 4: Points $y_1\underline{\phi}(x_1)$ and $y_2\underline{\phi}(x_2)$ along with the contours of the regularization term $\|\underline{\theta}\|$.