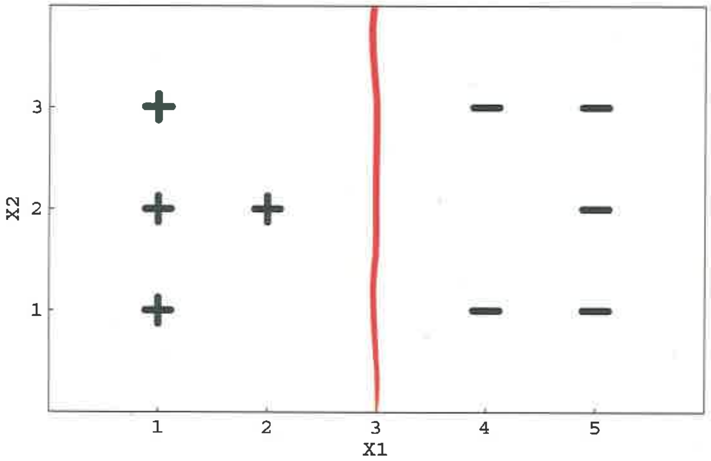
2 [40 points] Support Vector Machines

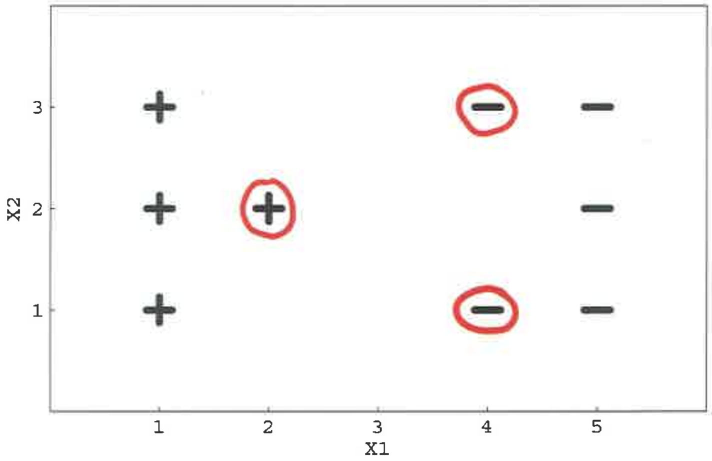
1. [4 points] Suppose we are using a linear SVM (i.e., no kernel), with some large C value, and are given the following data set.



Draw the decision boundary of linear SVM. Give a brief explanation.

Solution. Because of the large C value, the decision boundary will classify all of the examples correctly. Furthermore, among separators that classify the examples correctly, it will have the largest margin (distance to closest point).

2. [8 points] In the following image, circle the points such that after removing that point (example) from the training set and retraining SVM, we would get a different decision boundary than training on the full sample.



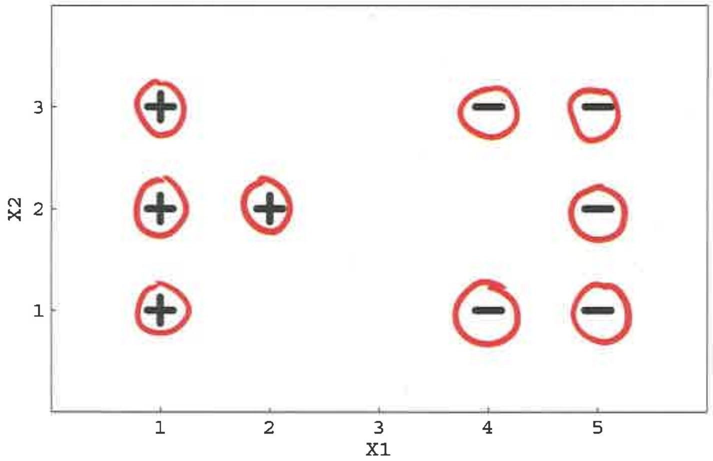
You do not need to provide a formal proof, but give a one or two sentence explanation.

Solution. These examples are the support vectors; all of the other examples are such that their corresponding constraints are not tight in the optimization problem, so removing them will not create a solution with smaller objective function value (norm of w). These three examples are positioned such that removing any one of them introduces slack in the constraints, allowing for a solution with a smaller objective function value and with a different third support vector; in this case, because each of these new (replacement) support vectors is not close to the old separator, the decision boundary shifts to make its distance to that example equal to the others.

3. [8 points] Suppose instead of SVM, we use regularized logistic regression to learn the classifier. That is,

$$(w, b) = \arg \min_{w\in\mathbb{R}^2, b\in\mathbb{R}} \frac{\|w\|^2}{2} - \sum_i \mathbb{1}[y_i = 0] \ln \frac{1}{1 + e^{(w\cdot x_i + b)}} + \mathbb{1}[y_i = 1] \ln \frac{e^{(w\cdot x_i+b)}}{1 + e^{(w\cdot x_i + b)}}.$$

In the following image, circle the points such that after removing that point (example) from the training set and running regularized logistic regression, we would get a different decision boundary than training with regularized logistic regression on the full sample.



You do not need to provide a formal proof, but give a one or two sentence explanation.

Solution. Because of the regularization, the weights will not diverge to infinity, and thus the probabilities at the solution are not at 0 and 1. Because of this, *every* example contributes to the loss function, and thus has an influence on the solution.

4. [8 points] Suppose we have a kernel $K(\cdot, \cdot)$, such that there is an implicit high-dimensional feature map $\phi : \mathbb{R}^d \to \mathbb{R}^D$ that satisfies $\forall x_i, x_j \in \mathbb{R}^d$, $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$, where $\phi(x_i) \cdot \phi(x_j) = \sum_{l=1}^{D} \phi(x_i)^l \phi(x_j)^l$ is the dot product in the D-dimensional space, and $\phi(x_i)^l$ is the l^{th} element/feature in the D-dimensional space.

Show how to calculate the Euclidean distance in the D-dimensional space

$$\|\phi(x_i) - \phi(x_j)\| = \sqrt{\sum_{l=1}^{D} (\phi(x_i)^l - \phi(x_j)^l)^2}$$

without explicitly calculating the values in the D-dimensional space. For this question, please ~~provide a formal~~ proof. in K(,)...

Hint: Try converting the Euclidean distance into a set of inner products.

Solution.

$$\begin{aligned}
\|\phi(x_i) - \phi(x_j)\| &= \sqrt{\sum_{l=1}^{D} (\phi(x_i)^l - \phi(x_j)^l)^2} \\
&= \sqrt{\sum_{l=1}^{D} (\phi(x_i)^l)^2 + (\phi(x_j)^l)^2 - 2\phi(x_i)^l \phi(x_j)^l} \\
&= \sqrt{\left(\sum_{l=1}^{D} (\phi(x_i)^l)^2\right) + \left(\sum_{l=1}^{D} (\phi(x_j)^l)^2\right) - \left(\sum_{l=1}^{D} 2\phi(x_i)^l \phi(x_j)^l\right)} \\
&= \sqrt{\phi(x_i) \cdot \phi(x_i) + \phi(x_j) \cdot \phi(x_j) - 2\phi(x_i) \cdot \phi(x_j)} \\
&= \sqrt{K(x_i, x_i) + K(x_j, x_j) - 2K(x_i, x_j)}.
\end{aligned}$$

5. [6 points] Assume that we use the RBF kernel function $K(x_i, x_j) = \exp(-\frac{1}{2}\|x_i - x_j\|^2)$. Also assume the same notation as in the last question. Prove that for any two input examples x_i and x_j, the squared Euclidean distance of their corresponding points in the high-dimensional space $\mathbb{R}^D$ is less than 2, i.e., prove that $\|\phi(x_i) - \phi(x_j)\|^2 < 2$.

Solution. This inequality directly follows from the result from the last question.

6. [6 points] Assume that we use the RBF kernel function, and the same notation as before. Consider running One Nearest Neighbor with Euclidean distance in both the input space $\mathbb{R}^d$ and the high-dimensional space $\mathbb{R}^D$. Is it possible that One Nearest Neighbor classifier achieves better classification performance in the high-dimensional space than in the original input space? Why?

Solution. No.

$$k(u, v) = \exp\left(-\frac{\|u-v\|^2}{2\sigma^2}\right)$$

$\|u, v\|^2$ smallest $\|u, v'\|^2$