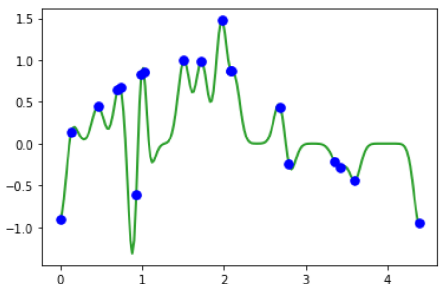
1 Kernel Regression

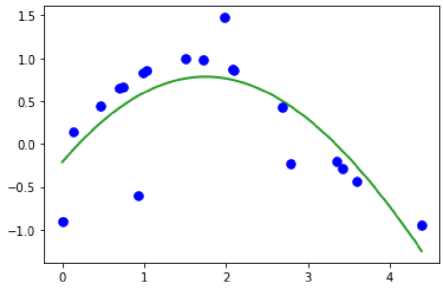
1.1 Conceptual Recap

a) Is kernel regression a parametric or nonparametric model? Explain.

It is a nonparametric model. The kernel is applied to each data point, so the number of parameters increase as the number of data points increase.

b) Consider the RBF kernel, $k(x, x^{(i)}) = e^{-\gamma \|x - x^{(i)}\|_2^2}$. Match the models below with their corresponding γ values (0.01 or 100). What is the effect of γ on overfitting?





The left model corresponds to $\gamma = 100$, while the right corresponds to $\gamma = 0.01$. When γ grows too large, the model is likely to overfit.

1.2 Kernelize it!

Recall the process for kernel regression:

Step 1: Compute $\boldsymbol{\alpha = (K + \lambda \mathbb{I})^{-1} y}$ where: $\boldsymbol{K_{ij} = k(x^{(i)}, x^{(j)})}$ and $\mathbf{k}$ is your kernel.

Step 2: Given a new point $\mathbf{x}$, predict $\boldsymbol{\hat{y} = \sum_{i=1}^{N} \alpha_i k(x, x^{(i)})}$

Monday's lecture introduced the box kernel and RBF kernel. For this question, consider the box kernel:

$$k(x, x^{(i)}) = \begin{cases} 1 & \|x - x^{(i)}\|_2^2 \leq \frac{1}{2} \\ 0 & otherwise \end{cases}$$

You are given the following 1D data points (x,y): (3,4), (3.7,1), (4.2,-2).

c) Find K and calculate α. For simplicity, let $\lambda = 0$. You can use an online inverse matrix calculator.

K (3x3): $\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$, where K^{-1} (3x3): $\begin{bmatrix} 0 & 1 & -1 \\ 1 & -1 & 1 \\ -1 & 1 & 0 \end{bmatrix}$

y (3x1): [4, 1, -2]

With $\gamma = 0$, we get α (3x1): $K^{-1}y = \begin{bmatrix} 0 & 1 & -1 \\ 1 & -1 & 1 \\ -1 & 1 & 0 \end{bmatrix}$ [4, 1, -2] = [3, 1, -3]

d) Predict $\hat{\mathbf{y}}$ for $\mathbf{x} = 3.4$.

$\hat{y} = \sum_{i=1}^{3} \alpha_i k(3.4, x^{(i)})$

$= \alpha_1 k(3.4, 3) + \alpha_2 k(3.4, 3.7) + \alpha_3 k(3.4, 4.2)$

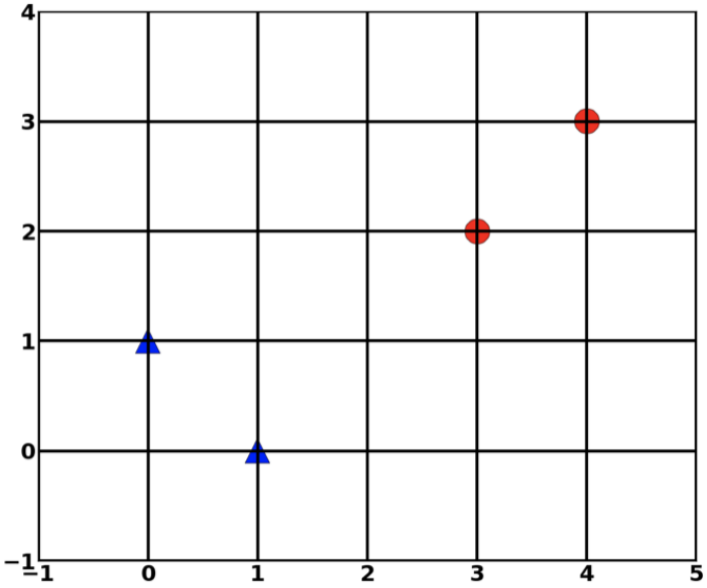
= 3*1 + 1*1 + (-3)*0

= 4

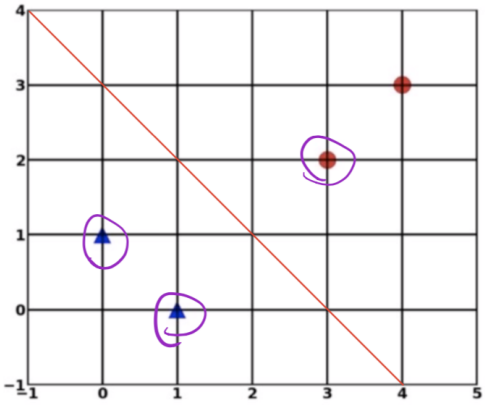
2 Support Vector Machines

Assume we are given dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{+1, -1\}$.
In SVM the goal is to find some hyperplane which separates the positive from the negative examples, such that the margin (the minimum distance from the decision boundary to the training points) is maximized. Let the equation for the hyperplane be $\mathbf{w}^T\mathbf{x} + b = 0$.

(a) You are presented with the following set of data (triangle = +1, circle = −1):



Find the equation of the hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$ that would be used by an SVM classifier. Which points are support vectors?



Support vectors are circled.

(b) Let's try to measure the width of the SVM slab (we assume that it was fitted to linearly separable data). We can do this by measuring the distance from one of the support vectors, say $\mathbf{x}_+$, to the plane

$\mathbf{w}^T\mathbf{x} + b = 0$ Since the equation of the plane is $\mathbf{w}^T\mathbf{x} + b = 0$ and since $c\mathbf{w}^T\mathbf{x} + b = 0$ defines the same plane, we have the freedom to choose the normalization of w. Let us choose normalization such that $\mathbf{w}^T\mathbf{x}_+ + b = +1$ and $\mathbf{w}^T\mathbf{x}_- + b = -1$, for the positive and negative support vectors respectively. Show that the width of an SVM slab with linearly separable data is $\frac{2}{||\mathbf{w}||}$.

Let d_+ denote the margin from $\mathbf{x}_+$ to the plane $\mathbf{w}^T\mathbf{x} + b = 0$. Let d_- denote the margin from $\mathbf{x}_-$ to the plane. The width then can be written as:

$$
\begin{aligned}
width &= d_+ + d_- \\
&= \frac{\mathbf{w}^T\mathbf{x}_+}{||\mathbf{w}||} + b - \frac{\mathbf{w}^T\mathbf{x}_-}{||\mathbf{w}||} - b && \text{(Project } x \text{ to } w\text{, normalize and translate)} \\
&= \frac{\mathbf{w}^T\mathbf{x}_+}{||\mathbf{w}||} - \frac{\mathbf{w}^T\mathbf{x}_-}{||\mathbf{w}||} \\
&= \frac{1-b}{||\mathbf{w}||} - \frac{-1-b}{||\mathbf{w}||} && \text{(Since both } x_+ \text{ and } x_- \text{ are support vectors)} \\
&= \frac{1}{||\mathbf{w}||}((1-b) - (-1-b)) \\
&= \frac{2}{||\mathbf{w}||}
\end{aligned}
$$

Thus we've proved that the width is $\frac{2}{||\mathbf{w}||}$

(c) Write SVM as an optimization problem. Conclude that maximizing the margin is equivalent to minimizing $||\mathbf{w}||$. In other words, write the SVM as some $max_{w,b}\frac{2}{||\mathbf{w}||}$ with certain constraints and show that the maximization problem implies $min_{w,b}||\mathbf{w}||$ on some constraints.

The main idea for SVM is to find a linear separator with a maximum margin. Using what we've proved in b), we can write SVM as follows:

$$
\begin{aligned}
&max_{w,b}\frac{2}{||\mathbf{w}||} \\
s.t. &y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 \qquad \forall i
\end{aligned}
$$

By inverting the objective function, we get the following equivalent optimization problem:

$$
\begin{aligned}
&min_{w,b}\frac{||\mathbf{w}||}{2} \\
s.t. &y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 \qquad \forall i
\end{aligned}
$$

Note that in the constraint the $\mathbf{x}^{(i)}s$ are not only the support vectors, but every data point.

(d) Will moving points which are not support vectors further away from the decision boundary effect the SVM's the w and b returned by the SVM optimization?

No. Since SVM only compute w and b based on support vectors, which are positive and negative samples that lie on the margin.