1 Prediction with SVM

Recall our formulation of the SVM dual:

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y^{(i)} y^{(j)} x^{(i)T} x^{(j)}$$

such that

$$\alpha_i \geq 0, i = 1, ..., N$$

Suppose we get a new point x and we would like to predict its label (-1 or 1). How do we compute this prediction?

We compute $\hat{y} = sign(w^T x + b)$. But where do we get w and b? Recall from lecture that w came from the Lagrangian, giving us $w = \sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)}$, so w is a linear combination of the training points multiplied by their labels. To find b, note that from lecture we saw that each α_i corresponded to a training point, and that the support vector were all the points where $\alpha_i > 0$.

$\alpha_i > 0$ implies that the ith constraint is active and tight, meaning that we have $y^{(i)}(w^T x^{(i)} + b) = 1$. Solving algebraically, we get $w^T x^{(i)} + b = y^{(i)}$ since the label is 1 or -1, so $b = y^{(i)} - w^T x^{(i)}$. So, we can determine b directly from any constraint corresponding to a support vector.

1.1 Geometric Intuition for LaGrange Multipliers

Solve the following constrained optimization problem:

$$\min_{x,y} f(x, y) = x^2 + y^2$$

Subject to

$$g(x, y) = 0$$

where

$$g(x, y) = x + y - 1$$

We use the method of LaGrange Multipliers and explain some intuition for why it works at the end.

We introduce the LaGrangian $L(x, y, \lambda) = f(x, y) - \lambda g(x, y)$

The gradient is then $\nabla L(x, y, \lambda) = \nabla f(x, y) - \nabla \lambda g(x, y)$

We then set equal to 0 and solve the system:

$$\frac{\partial L}{\partial x} = 0$$
$$\frac{\partial L}{\partial y} = 0$$
$$\frac{\partial L}{\partial \lambda} = 0$$

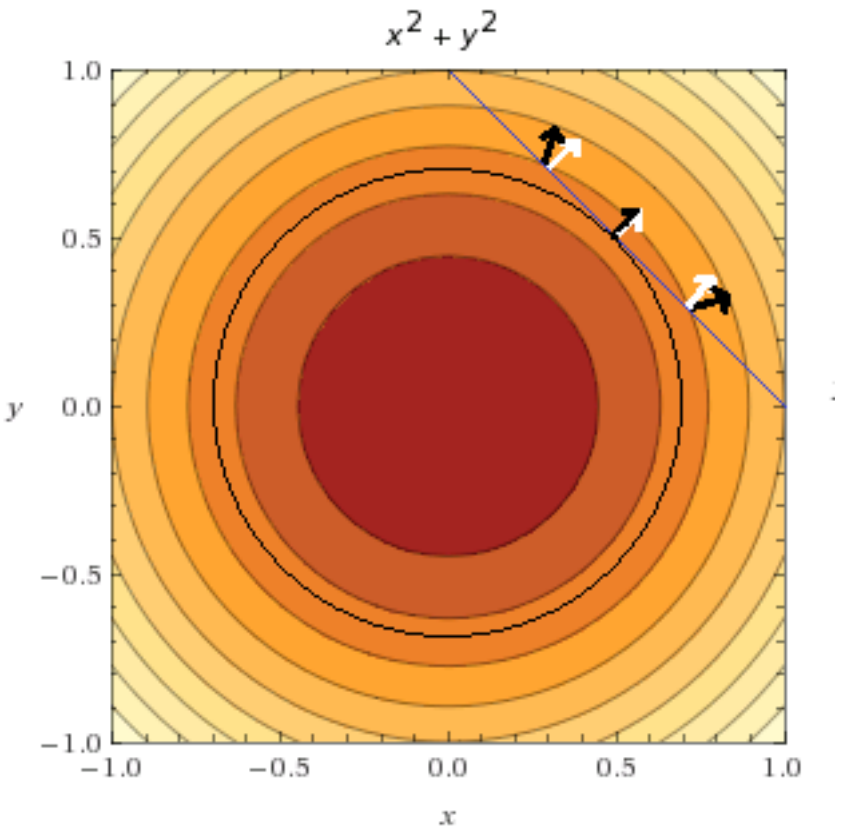
This then give us the system:

$$2x - \lambda = 0$$
$$2y - \lambda = 0$$
$$x + y = 1$$

Solving the resulting system gives us the optimum at $x = \frac{1}{2}, y = \frac{1}{2}$

The main observation for why LaGrange multipliers work is to see that the minimum of $f(x, y)$ under the constraint $g(x, y)$ occurs when their gradients are parallel.

Geometrically, we have the following contour plot:



The circles are the contours of the objective function $f(x, y) = x^2 + y^2$ with darker colors signalling smaller values. The line in the top right is the constraint $g(x, y) = y = 1 - x$. Black arrows signal the gradient of $f(x, y)$ and white arrows signal the gradient of $g(x, y)$. The claim is that the middle point, where the gradients are parallel (in the same direction in this case), is the minimum of the objective under the constraint.

To see this, suppose you start at the leftmost point (on the line labelled with the 3 points and gradients) Say that the objective function evaluated at this point gives us some value v.

You would like to go left, but then you see that the gradient of the objective function is pointing towards the left (remember it is always pointing towards greater values) so it might not be a good idea. To be sure, you try and go to the left. You find out that the objective function returns some value $x > v$, so it is increasing which is not okay. Going to the right, this time, the objective function returns smaller values, it is good you are going in the correct direction. So you continue like that until you reach the center point where the two

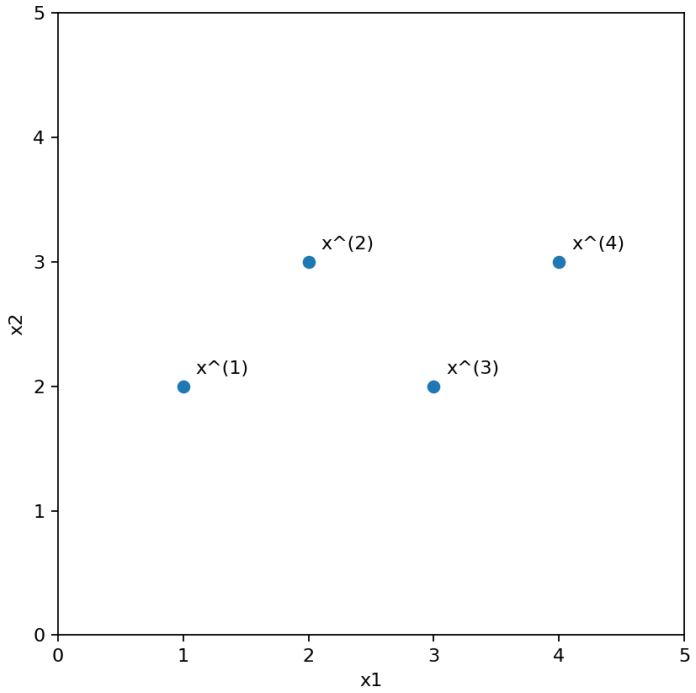
arrows are parallel. But then, once you move a little bit to the right you find out that the objective function is increasing again. As you can not move right or left anymore without increasing the objective function, you conclude this point is the minimum.

Finally, how does λ come into play? We know that the minimum occurs where the gradients are in the same direction. But their magnitudes may be different. Thus, λ is the factor where $\nabla f(x, y) = \lambda \nabla g(x, y)$ and $g(x, y) = 0$.

For a more in-depth look please see https://www.svm-tutorial.com/2016/09/duality-lagrange-multipliers. Credit is given to this website as well for the problem and explanation.

2 PCA: Basic Concepts

Consider dataset $\mathcal{D} = \{\mathbf{x}^{(1)} = \begin{bmatrix}1\\2\end{bmatrix}, \mathbf{x}^{(2)} = \begin{bmatrix}2\\3\end{bmatrix}, \mathbf{x}^{(3)} = \begin{bmatrix}3\\2\end{bmatrix}, \mathbf{x}^{(4)} = \begin{bmatrix}4\\3\end{bmatrix}\}$. A visualization of the dataset is as below.



2.1 Centering Data

Centering is crucial for PCA. We must preprocess data so that all features have zero mean before applying PCA, i.e.

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^{(i)} = \vec{0}$$

Compute the centered dataset:

$\mathbf{x}^{(1)} =$ _____ $\mathbf{x}^{(2)} =$ _____

$\mathbf{x}^{(3)} =$ _____ $\mathbf{x}^{(4)} =$ _____

First, note that $\hat{E}[x_1] = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_1^{(i)} = 2.5$ and $\hat{E}[x_2] = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_2^{(i)} = 2.5$

$$\mathbf{x}^{(1)} = \begin{bmatrix} -1.5 \\ -0.5 \end{bmatrix} \quad \mathbf{x}^{(2)} = \begin{bmatrix} -0.5 \\ 0.5 \end{bmatrix}$$

$$\mathbf{x}^{(3)} = \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix} \quad \mathbf{x}^{(4)} = \begin{bmatrix} 1.5 \\ 0.5 \end{bmatrix}$$

2.2 Unit vector

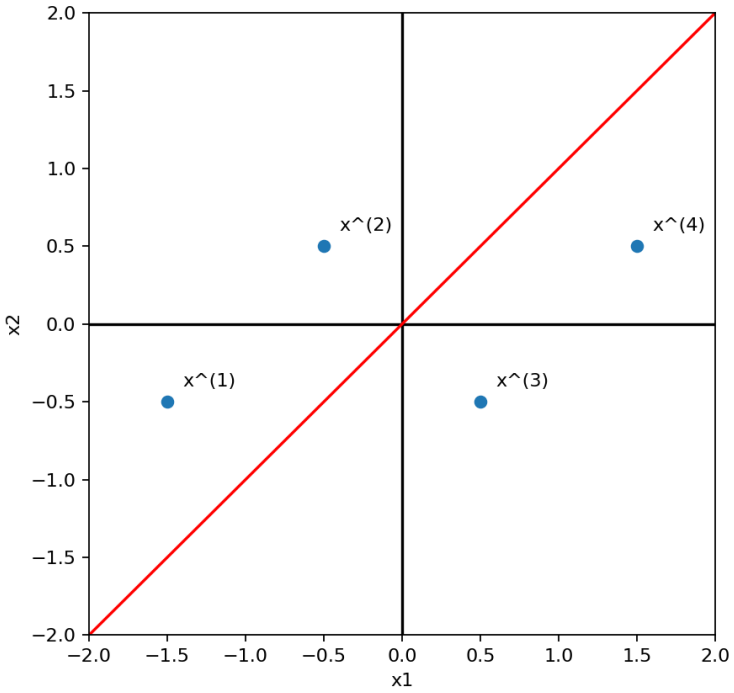
In order to easily compute the projected coordinates of data, we need to make the projected directions unit vectors. Suppose we want to project our data onto the vector $\mathbf{v} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Normalize $\mathbf{v}$ to be a unit vector.

$\mathbf{v} =$ _____

$$\mathbf{v} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

2.3 Project Data

The centered data should now look like the following:



Suppose we want to project the centered data onto $\mathbf{v}$, where $\mathbf{v}$ goes through the origin.

Compute the magnitude of the projections, i.e. compute $z^{(i)} = \mathbf{v}^T\mathbf{x}^{(i)}, \forall 1 \leq i \leq N$.

$z^{(1)} =$ _____ $z^{(2)} =$ _____

$z^{(3)} =$ _____ $z^{(4)} =$ _____

$$z^{(1)} = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}] \begin{bmatrix} -1.5 \\ -0.5 \end{bmatrix} = \frac{-3}{2\sqrt{2}} - \frac{1}{2\sqrt{2}} = -\frac{4}{2\sqrt{2}} = -\sqrt{2}$$
$$z^{(2)} = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}] \begin{bmatrix} -0.5 \\ 0.5 \end{bmatrix} = \frac{-1}{2\sqrt{2}} + \frac{1}{2\sqrt{2}} = 0$$
$$z^{(3)} = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}] \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix} = \frac{-1}{2\sqrt{2}} + \frac{1}{2\sqrt{2}} = 0$$
$$z^{(4)} = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}] \begin{bmatrix} 1.5 \\ 0.5 \end{bmatrix} = \frac{3}{2\sqrt{2}} + \frac{1}{2\sqrt{2}} = \frac{4}{2\sqrt{2}} = \sqrt{2}$$

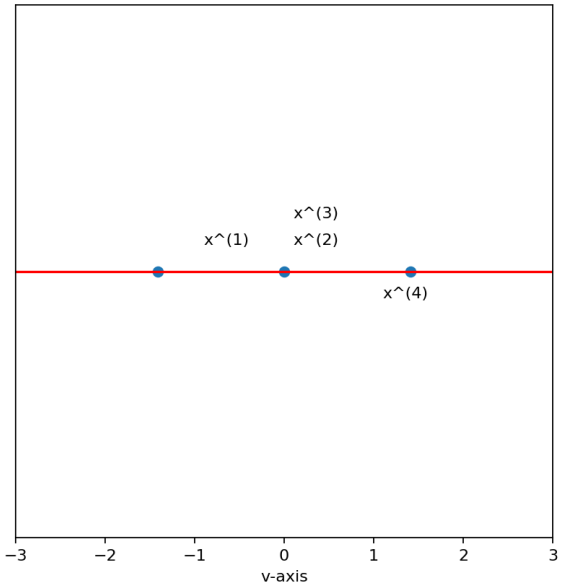
Let $\mathbf{x}^{(i)'}$ be the projected point of $\mathbf{x}^{(i)}$. Note that $\mathbf{x}^{(i)'} = \mathbf{v}^T\mathbf{x}^{(i)}\mathbf{v} = z^{(i)}\mathbf{v}$. Compute the projected coordinates:

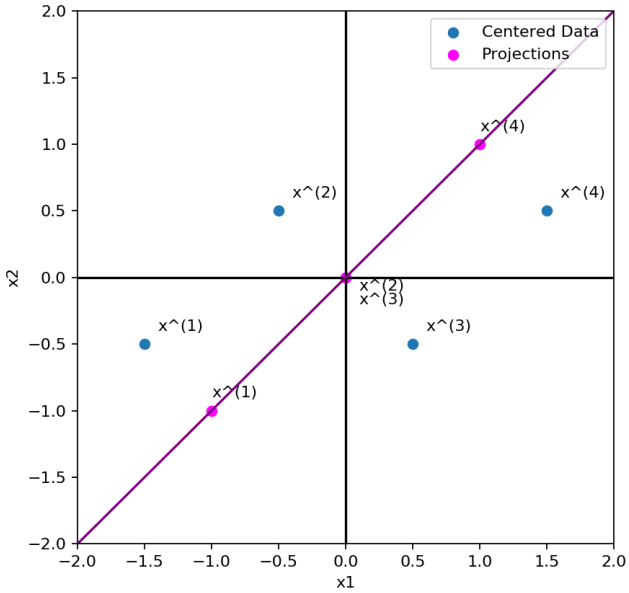
$\mathbf{x}^{(1)'} =$ _____ $\mathbf{x}^{(2)'} =$ _____

$\mathbf{x}^{(3)'} =$ _____ $\mathbf{x}^{(4)'} =$ _____

$$\mathbf{x}^{(1)'} = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \qquad \mathbf{x}^{(2)'} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$
$$\mathbf{x}^{(3)'} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \mathbf{x}^{(4)'} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Below is a visualization of the projections:





2.4 Reconstruction Error

One of the two goals of PCA is to find new directions to project our dataset onto such that it **minimizes the reconstruction error**, where the reconstruction error is defined as following:

$$\text{Reconstruction Error} = \sum_{i=1}^{N} \|\mathbf{x}^{(i)'} - \mathbf{x}^{(i)}\|_2^2$$

What is the reconstruction error in our case?

$$\text{Reconstruction Error} = \_\_\_\_\_$$

$$
\begin{aligned}
\text{Reconstruction Error} &= \sum_{i=1}^{N} \|\mathbf{x}^{(i)'} - \mathbf{x}^{(i)}\|_2^2 \\
&= \| \begin{bmatrix} -1 \\ -1 \end{bmatrix} - \begin{bmatrix} -1.5 \\ -0.5 \end{bmatrix} \|_2^2 + \| \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} -0.5 \\ 0.5 \end{bmatrix} \|_2^2 + \| \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix} \|_2^2 + \| \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 1.5 \\ 0.5 \end{bmatrix} \|_2^2 \\
&= \sqrt{(\frac{1}{2})^2 + (-\frac{1}{2})^2} + \sqrt{(\frac{1}{2})^2 + (-\frac{1}{2})^2} + \sqrt{(-\frac{1}{2})^2 + (\frac{1}{2})^2} + \sqrt{(-\frac{1}{2})^2 + (\frac{1}{2})^2} \\
&= \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} \\
&= \frac{4}{\sqrt{2}} = 2\sqrt{2}
\end{aligned}
$$

2.5 Variance of Projected Data

Another goal is to find new directions to project our dataset onto such that it **maximizes the variance of the projections**, where the variance of projections is defined as following:

$$
\begin{aligned}
\text{variance of projection} &= \sum_{i=1}^{N} (z^{(i)} - \hat{E}[z])^2 \\
&= \sum_{i=1}^{N} (\mathbf{v}^T \mathbf{x}^{(i)} - \frac{1}{N} \sum_{j=1}^{N} \{\mathbf{v}^T \mathbf{x}^{(j)})^2 \\
&= \sum_{i=1}^{N} (\mathbf{v}^T \mathbf{x}^{(i)} - \mathbf{v}^T (\frac{1}{N} \sum_{j=1}^{N} \mathbf{x}^{(j)}))^2 \\
&= \sum_{i=1}^{N} (\mathbf{v}^T \mathbf{x}^{(i)} - \mathbf{v}^T \vec{0})^2 \\
&= \sum_{i=1}^{N} (\mathbf{v}^T \mathbf{x}^{(i)})^2
\end{aligned}
$$

What is the variance of the projections?

$$\text{variance} = \_\_\_\_\_$$

$$
\begin{aligned}
\text{variance} &= \sum_{i=1}^{N} (z^{(i)})^2 \\
&= (-\sqrt{2})^2 + 0^2 + 0^2 + (\sqrt{2})^2 \\
&= 2 + 0 + 0 + 2 \\
&= 4
\end{aligned}
$$