Machine Learning Notes

Prerequisites, Regression and Classification

Raghav B. Venkataramaiyer

Feb '25

1 Setup

1. Given is a set of paired observations $\mathcal{D}$ (aka evidence), where targets $y \in \mathbb{R}$ are paired with (d dimensional) features $\mathbf{x} \in \mathbb{R}^d$.
2. We propose a mathematical model (typically a family of functions) $\mathcal{F}_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}$ parameterised by $\boldsymbol{\theta}$
3. So that $y \approx \mathcal{F}_{\boldsymbol{\theta}_*}(\mathbf{x})$. Here y are referred to as targets, $\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x})$ are referred to as predictions; so that predictions approximate the targets, under optimal set of learnt parameters, $\boldsymbol{\theta}_*$.
4. We express this formally as:
   Find $\boldsymbol{\theta} = \boldsymbol{\theta}_*$ in order to

$$\underset{\boldsymbol{\theta}}{\text{minimise}} \quad \mathop{\mathbb{E}}_{y,\mathbf{x}\sim\mathcal{D}} [\Delta(y, \mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}))]$$

   where, Δ is the notion of distance between predictions $\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x})$ and targets y.

2 Linear Regression

2.1 In 2D

$$y \approx \mathcal{F}_{w,b}(x) = wx + b$$
$$\Delta\left(y, \mathcal{F}_{w,b}(x)\right) = \frac{1}{2}\left(y - \mathcal{F}_{w,b}(x)\right)^2$$

The objective is to find $w = w_*$, $b = b_*$ in order to

$$\underset{w,b}{\text{minimise}} \quad \mathop{\mathbb{E}}_{y,x\sim\mathcal{D}} \left[\frac{1}{2}\left(y - \mathcal{F}_{w,b}(x)\right)^2\right]$$

The analytical solution yields,

$$\begin{aligned}
w_* &= \frac{\text{coVar}(x, y)}{\text{Var}(x)} \\
b_* &= \mathbb{E}[y] - w_*\mathbb{E}[x] \\
\text{coVar}(x, y) &= \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] \\
\text{Var}(x) &= \mathbb{E}[x^2] - \mathbb{E}^2[x]
\end{aligned}$$

2.2 In Higher Dimensions

$$\begin{aligned}
y \approx \mathcal{F}_{\mathbf{w}}(\mathbf{x}) &= \mathbf{w}^\top\mathbf{x} = \mathbf{x}^\top\mathbf{w} \\
&= w_0 + w_1x_1 + \cdots + w_dx_d \qquad (x_0 = 1) \\
\Delta\left(y, \mathcal{F}_{\mathbf{w}}(\mathbf{x})\right) &= \frac{1}{2}\left(y - \mathcal{F}_{\mathbf{w}}(\mathbf{x})\right)^2 \\
&= \frac{1}{2}\left(y - \mathbf{x}^\top\mathbf{w}\right)^2
\end{aligned}$$

The objective is to find $\mathbf{w} = \mathbf{w}_*$ in order to

$$\begin{aligned}
&\underset{\mathbf{w}}{\text{minimise}} \quad \mathop{\mathbb{E}}_{y,\mathbf{x}\sim\mathcal{D}} \left[\frac{1}{2}\left(y - \mathbf{x}^\top\mathbf{w}\right)^2\right] \\
\text{or,} \quad &\underset{\mathbf{w}}{\text{minimise}} \quad \frac{1}{2}(\mathbf{y} - X\mathbf{w})^\top(\mathbf{y} - X\mathbf{w}) \\
\text{where,} \quad &\mathbf{y} \equiv \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \quad X \equiv \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix}
\end{aligned}$$

The analytical solution yields,

$$\mathbf{w}_* = (X^\top X)^{-1}X^\top\mathbf{y}$$

2.3 Implementation

2.3.1 In Spreadsheet

This (Google Sheet) will help understand and practice computing the solution manually for the case in 2D.

2.3.2 In Code

This (Gist) is a reference python implementation of the analytical solution.

3 Logistic Regression

(Binary Classification)

$$
\begin{aligned}
y \approx \widetilde{y} = \mathcal{F}_{\mathbf{w}}(\mathbf{x}) &= \sigma(\mathbf{x}^\top \mathbf{w}) \\
\Delta\left(y, \widetilde{y}\right) &= -\left(y \ln \widetilde{y} + (1-y)\ln(1-\widetilde{y})\right) \\
\frac{\partial \Delta(y, \widetilde{y})}{\partial \mathbf{w}} &= -(y - \widetilde{y})\mathbf{x}
\end{aligned}
$$

The objective is to find $\mathbf{w} = \mathbf{w}_*$ in order to

$$
\underset{\mathbf{w}}{\text{minimise}} \quad \mathcal{L}(\mathbf{w}) = \mathop{\mathbb{E}}_{y,\mathbf{x}\sim\mathcal{D}}\left[\Delta\left(y, \widetilde{y}\right)\right]
$$

Theres no analytical solution. But using gradient descent, we numerically hope to converge using iterative update,

$$
\begin{aligned}
\mathbf{w} &\leftarrow \mathbf{w} - \lambda \frac{\partial \mathcal{L}}{\partial \mathbf{w}} \\
&= \mathbf{w} + \lambda \mathop{\mathbb{E}}_{y,\mathbf{x}\sim\mathcal{D}}\left[(y - \widetilde{y})\mathbf{x}\right]
\end{aligned}
$$

Corrigendum 2025-02-22

1. Typo in Binary Cross Entropy Loss. Earlier mentioned: $\Delta = y \ln \widetilde{y} + \cdots$. This has been corrected now.
2. By cascading effect, $\partial\Delta/\partial\mathbf{w}$, and the update step for gradient descent algorithm also have been corrected.

4 Support Vector Machine

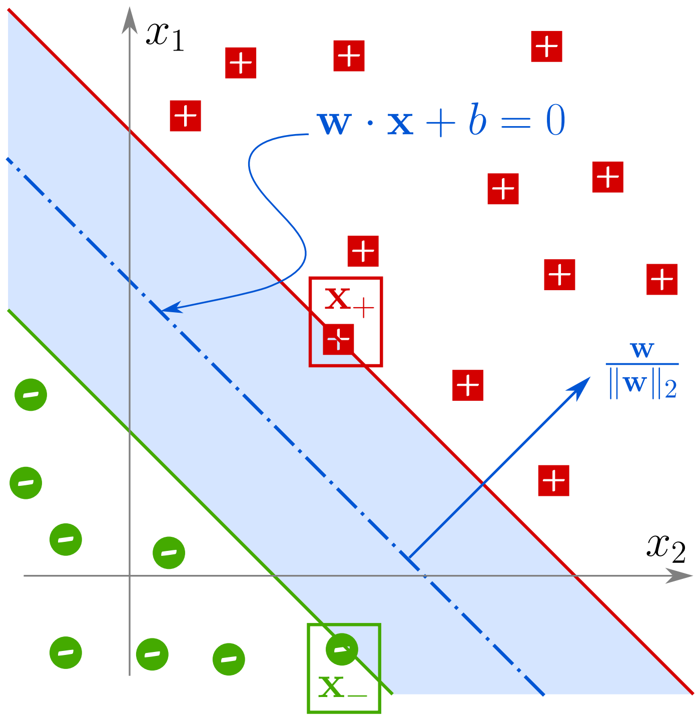

Figure 1: SVM Theory Illustration

1. Given a dataset $\mathcal{D}$ with paired samples $(y, \mathbf{x}); y \in \{+1, -1\}$ so that positive samples are labeled $y = +1$, and similarly negative samples as $y = -1$.

2. To evaluate for a simple case, lets assume that the positive and negative samples are comfortably separable through **a hyperplane**. In case of 2D data ($\mathbf{x} \in \mathbb{R}^2$), it would follow from the assumption that there exists a straight line with a finite margin, called **gutter space** such that,

   (a) There are no samples in the gutter space;

   (b) Positive samples lie on one side of the hyperplane; and

   (c) Negative samples lie on the other side.

3. Our aim is to find the straight line that maximises the gutter space.

4. Let the separating hyperplane (straight line in case of 2D data) be given as,

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \tag{1}$$

Geometrically speaking, $\mathbf{w}$ is a vector normal to the separating hyperplane. And the unit vector in the same direction is given as $\mathbf{w}/\|\mathbf{w}\|_2$. Where $\|\mathbf{w}\|_2$ is called the Frobenius Norm and $\|\mathbf{w}\|_2^2 = w_1^2 + \cdots + w_d^2$. This is the same as the understanding of magnitude of the vector in Euclidean space.

5. The hyperplane separates the space such that
One side of it satisfies $\mathbf{w} \cdot \mathbf{x} + b < 0$; and
The other side satisfies $\mathbf{w} \cdot \mathbf{x} + b > 0$.

6. From the separability assumption, it follows,

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x} + b < 0 \quad \forall y = -1 \\ \mathbf{w} \cdot \mathbf{x} + b > 0 \quad \forall y = +1 \end{aligned}$$

7. From the margin assumption, without loss of generality, it follows that

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x} + b \leqslant -1 \quad \forall y = -1 \\ \mathbf{w} \cdot \mathbf{x} + b \geqslant 1 \quad \forall y = +1 \end{aligned}$$

8. In other words

$$y(\mathbf{w} \cdot \mathbf{x} + b) \geqslant 1 \tag{2}$$

9. For the points on the margin, denoted as $\mathbf{x}_+, \mathbf{x}_-$ in the adjoining image,

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x}_+ + b &= 1 \\ \mathbf{w} \cdot \mathbf{x}_- + b &= -1 \\ \mathbf{w} \cdot (\mathbf{x}_+ - \mathbf{x}_-) &= 2 \end{aligned} \tag{3}$$

10. The gutter width γ is given as the projection of vector $\mathbf{x}_+ - \mathbf{x}_-$ along the normal to the hyperplane. Or,

$$\begin{aligned}
\gamma &= \frac{\mathbf{w}}{\|\mathbf{w}\|_2} \cdot (\mathbf{x}_+ - \mathbf{x}_-) \\
&= \frac{\mathbf{w} \cdot (\mathbf{x}_+ - \mathbf{x}_-)}{\|\mathbf{w}\|_2} \\
\gamma &= \frac{2}{\|\mathbf{w}\|_2} \qquad (4)
\end{aligned}$$

Our aim is to maximise the gutter width γ, which would be the same as minimising $1/\gamma$, or $1/\gamma^2$, or $4/\gamma^2 = \|\mathbf{w}\|_2^2$.

4.1 Training

Formally speaking, we need to find the parameters $\mathbf{w}, b$ in order to

$$\begin{aligned}
\text{minimise} \quad & \|\mathbf{w}\|_2^2 \\
\text{subject to,} \quad & y(\mathbf{w} \cdot \mathbf{x} + b) \geqslant 1
\end{aligned}$$

4.2 Inference

For all unseen points, $\mathbf{x}$, the estimated label $\widehat{y}$ is given as,

$$\widehat{y} = \text{signum}(\mathbf{w} \cdot \mathbf{x} + b) \qquad (5)$$

4.3 Implementation

Check out this gist