1.Derivation of optimal w for a binary SVM

2.Derivation of optimal w's for a3-layer MLP

3.Define a loss-function for SVM

1. Derivation of optimal w for a binary SVM

給定D=(x1,y1),(x2,y2)….(xm,ym),yi∈-1,+1

SVM考慮基於訓練集 D 在樣本空間中找到一個劃分超平面(hiperplane)，將不同類別的樣本分開。

劃分超平面公式：

其中為法向量，決定了超平面的方向； b為位移項，決定了超平面與原點之間的距離。通常劃分超平面用來表示，因為其可被法向量 和位移 確定。

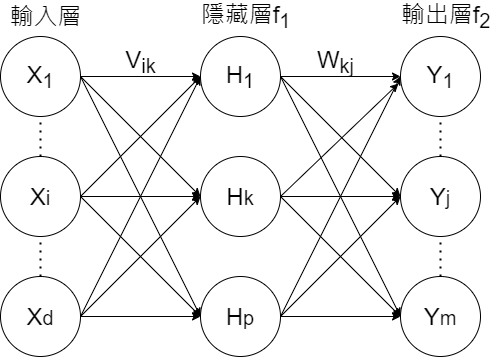
樣本空間中任意點 到超平面的距離可寫為

假設超平面 (w,b) 能將訓練樣本正確分類，對於 ,有以下式子：

其中距離超平面最近的幾個訓練樣本點使上式的等號成立，這幾個訓練樣本就被稱作支援向量（support vector），兩個異類支持向量到超平面的距離之和，也稱為間隔（margin），為，而要最佳化SVM，要做的就是讓他們之間的間隔最大，才能有效劃分資料，即讓最大，用數學公式來表示是這個樣子。

為了最大化間隔，僅需最大化，這等價於最小化 。於是上式我們可以重寫為

1. Derivation of optimal w's for a3-layer MLP



為第i筆資料的輸入值，其輸出值為

其目標的誤差為

所有樣本的誤差和當作目標函數

最佳化的目的就是讓「所有樣本的誤差均方和」越小越好，所以目標是

所以要找到最佳參數解(參數只有和)，最簡單的方式就是微分方程式等於0找解

但參數量多無法直接找到唯一解，所以要gradient descent找最佳解。

其中η為學習率(learning rate)

輸出到隱藏層(wkj)

chain rule:

第一項:

第二項

得

隱藏層到輸入層(vik)

chain rule:

整合起來

第一項:

第二項:

第三項

得

最後把n個樣本所有gradient加起來得到參數的update

1. Define a loss-function for SVM

損失函數

寫成算式就是

接下來要找到他的極值，我們要針對對L進行微分

接下來針對b對L進行微分

從上面的表達式，可得