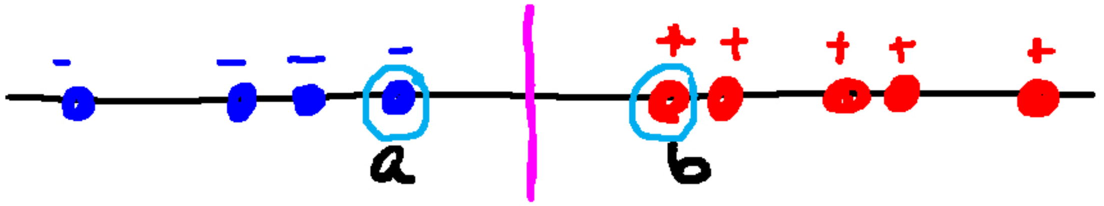
Support Vector Machines for Classification

Objectives

- Define margin for separable data
- Show support vector machines maximize margin
- Use hinge loss to define support vector machines for nonseparable data

Maximize margin for separable training data

Example:

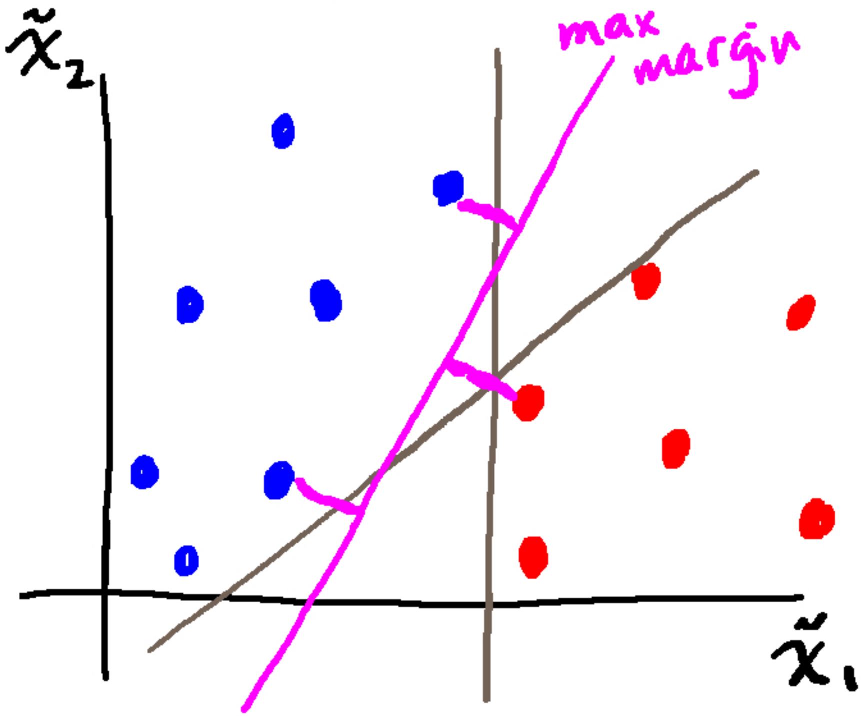


decision boundary?

margin: distance from boundary to nearest sample

max margin boundary: midpoint

only a, b matter



decision boundary?

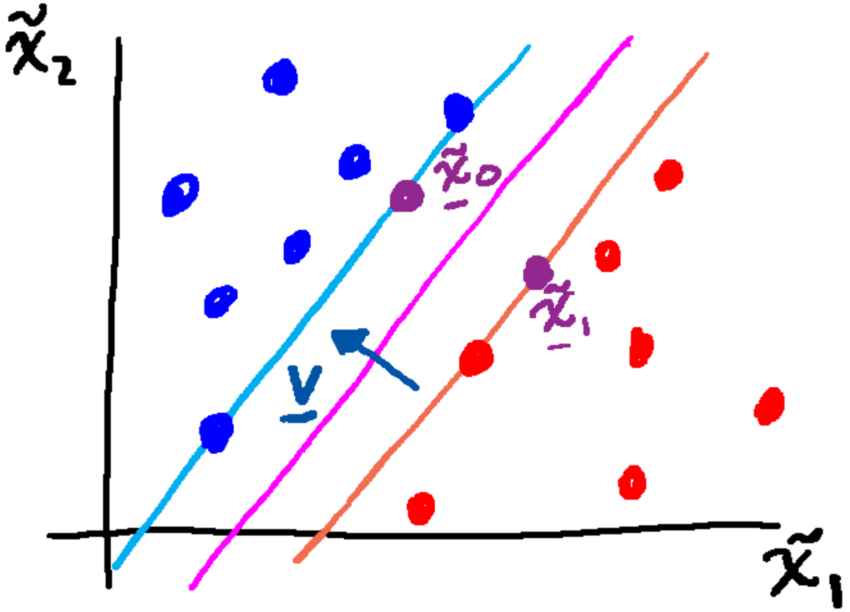
feature $\underline{x}^T = [\, \underline{\tilde{x}}^T \quad 1 \,]$

weights $\underline{w}^T = [\, \underline{\tilde{w}}^T \quad w_0 \,]$

decision $\hat{d} = \text{sign}(\underline{x}^T \underline{w})$

$$\hat{d} = \begin{cases} 1 & \underline{\tilde{x}}^T \underline{\tilde{w}} + w_0 > 0 \\ -1 & \underline{\tilde{x}}^T \underline{\tilde{w}} + w_0 < 0 \end{cases}$$

Margin is determined by $\|\tilde{\underline{w}}\|_2^{-1}$



label "-1": $\tilde{\underline{x}}^T\tilde{\underline{w}} + w_0 \leq -1$

label "+1": $\tilde{\underline{x}}^T\tilde{\underline{w}} + w_0 \geq 1$

boundary: $\tilde{\underline{x}}^T\tilde{\underline{w}} + w_0 = 0$

margin: ½ distance between / /
measure in direction $\underline{v}$

Unit normal to boundary plane: $\underline{v} = \dfrac{\tilde{\underline{w}}}{\|\tilde{\underline{w}}\|_2}$

Margin $m = \frac{1}{2}\|\tilde{\underline{x}}_1 - \tilde{\underline{x}}_0\|_2$ $\quad \tilde{\underline{x}}_1 = \tilde{\underline{x}}_0 + 2m\,\underline{v}$

$$1 = \tilde{\underline{x}}_1^T\tilde{\underline{w}} + w_0 = \tilde{\underline{x}}_0^T\tilde{\underline{w}} + 2m\frac{\tilde{\underline{w}}^T}{\|\tilde{\underline{w}}\|_2}\tilde{\underline{w}} + w_0$$

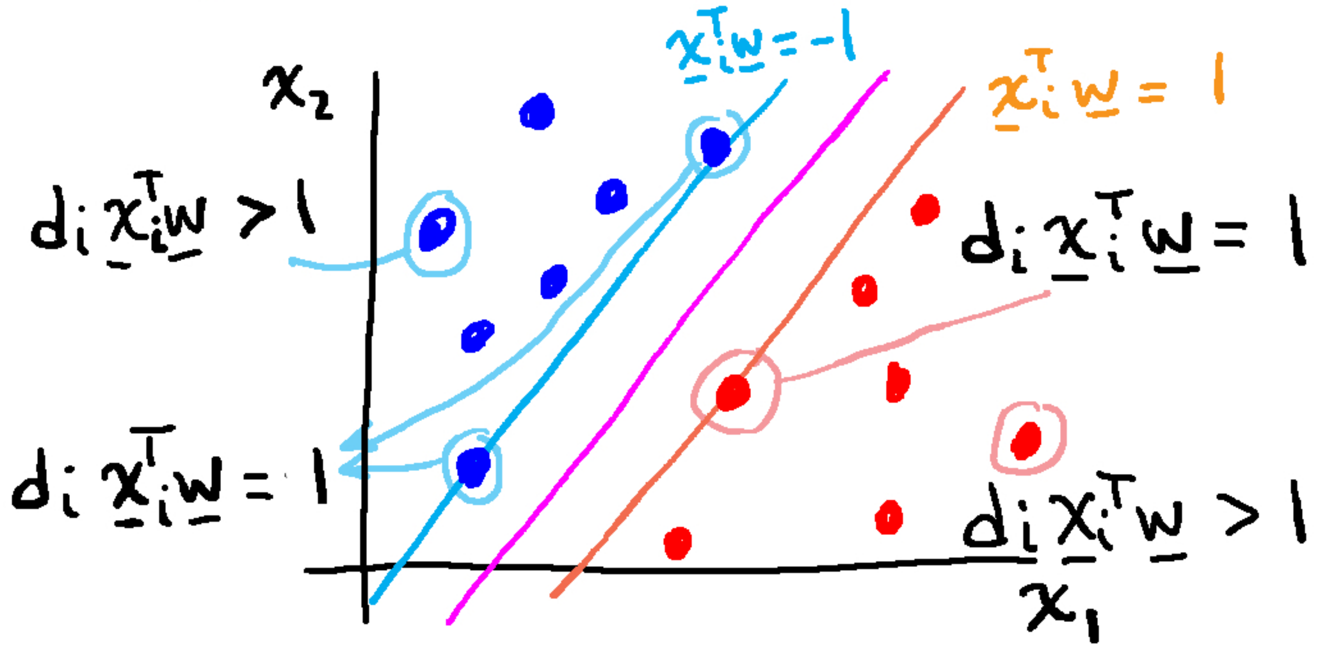
but $\tilde{\underline{x}}_0^T\tilde{\underline{w}} + w_0 = -1$

so $2 = 2m\|\tilde{\underline{w}}\|_2$

$$m = \|\tilde{\underline{w}}\|_2^{-1}$$

Support Vector Machine maximizes margin

Correct classification: $d_i(\tilde{\underline{x}}_i^T \tilde{\underline{w}} + w_0) \geq 1$



SVM:

$$\underbrace{\min_{\tilde{\underline{w}}, w_0} \|\tilde{\underline{w}}\|_2^2}_{\text{max margin}} \quad \text{s.t.} \quad \underbrace{d_i(\tilde{\underline{x}}_i^T \tilde{\underline{w}} + w_0) \geq 1 \quad i = 1, 2, \ldots, N}_{\text{perfect classification}}$$

unique solution

Boundary defined by $\underline{x}_i$ for which $d_i \underline{x}_i^T \underline{w} = 1$

Called Support Vectors

SVM for nonseparable data uses hinge loss

$$\min_{\underline{w}} \underbrace{\sum_{i=1}^{N} (1 - d_i \underline{x}_i^T \underline{w})_+}_{\text{loss}} + \lambda \underbrace{\|\underline{\tilde{w}}\|_2^2}_{\sim (\text{margin})^{-2}} \qquad (\ell_2 \text{ regularization})$$