CSDS 440: Machine Learning

Soumya Ray (he/him, sray@case.edu)

Olin 516

Office hours T, Th 11:15-11:45 or by appointment

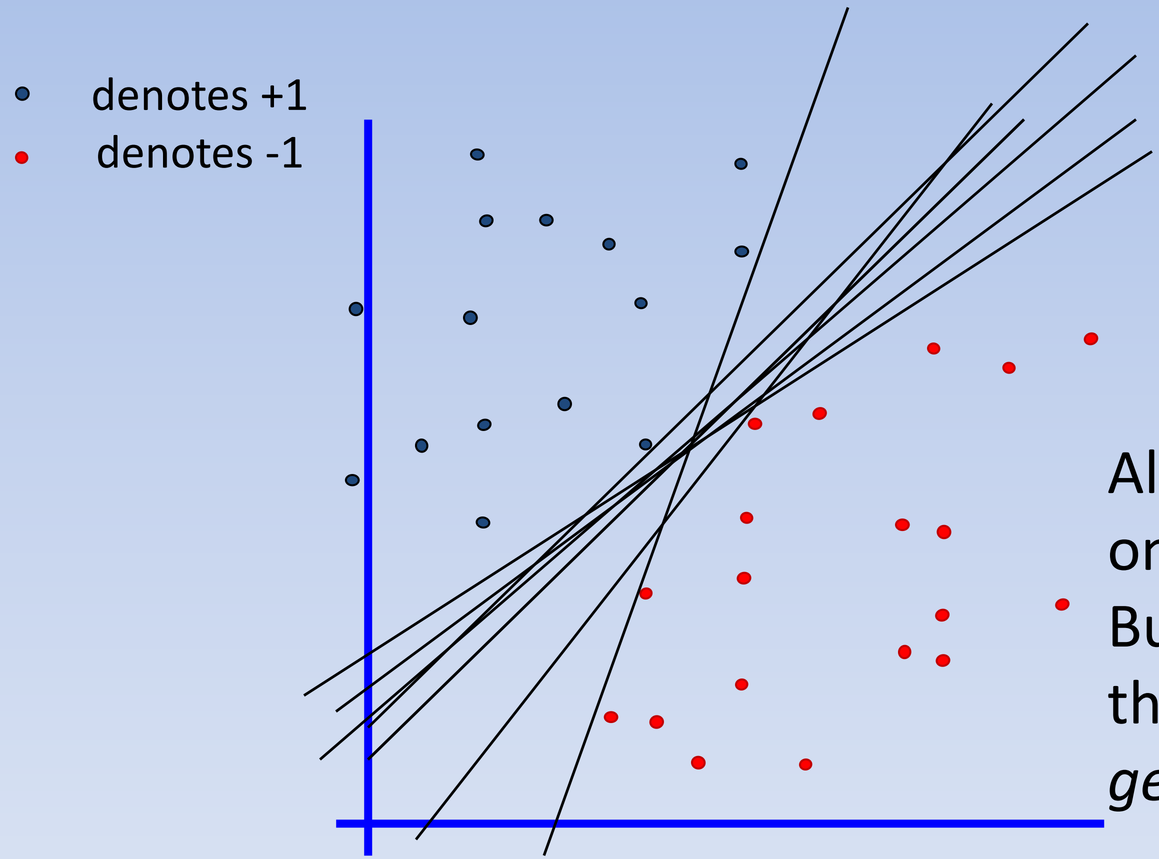
Recap

- Using the KL divergence, we can write the log likelihood of the training set approximately as (expectation of log(P(__ | __))+ D(___ || __). This holds when ___ = ___.
- This is an “e___ d___” architecture. The e___ is ____. The d___ is ____.
- To compute the first term we only need o__ sample. This is because ______.
- However, this is problematic because _____. To solve this, we sample from a noise distribution at the input layer, then multiply and add. This is called the r____ t____.
- Support vector machines combine three ideas: l___ d____, m____ and k____.
- An LD has the general form ______.

Today

- Support Vector Machines

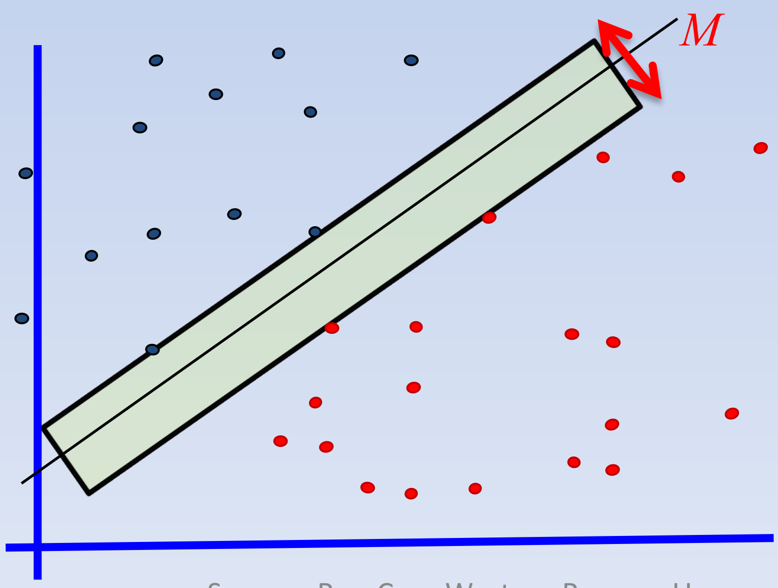
Find the Classifier



All are equally good on the training sample. But is there any one that we expect to *generalize* best?

Margin of a Classifier

- Imagine sliding any linear classifier parallel to itself
- The sum of the amounts we can move until we hit an (some) example(s) is the margin



Support Vector Machine

- The linear classifier with the maximum margin is called a support vector machine classifier

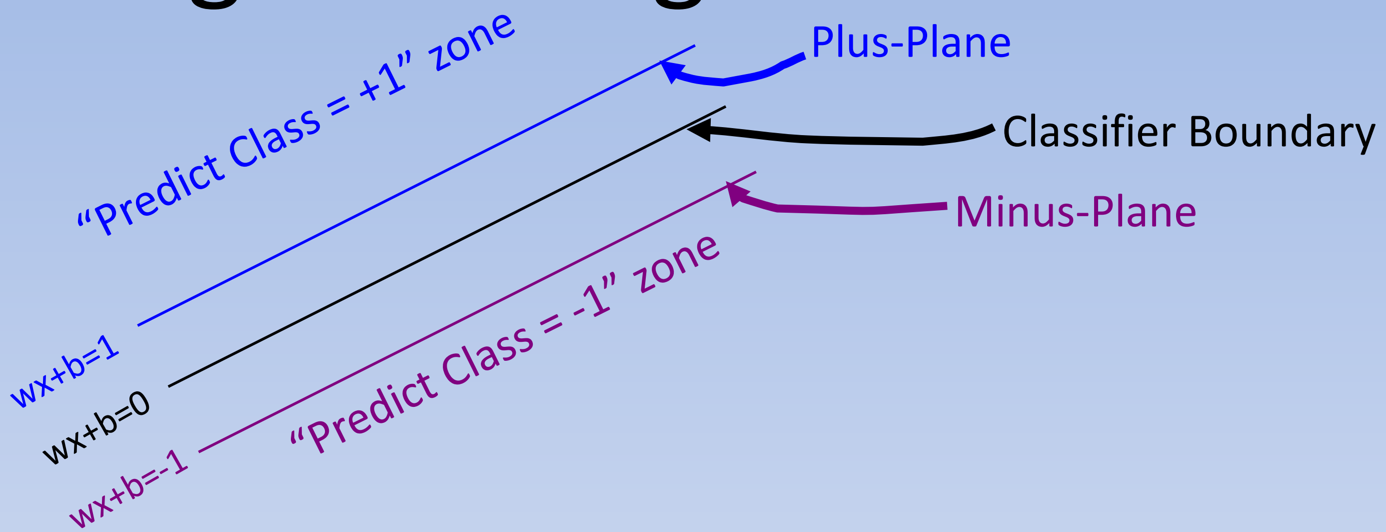
- If we are in the input feature space , i.e. $\boldsymbol{\varphi}(\mathbf{x})=\mathbf{x}$, this is called a linear SVM

- The examples that touch the margin boundaries are called the support vectors

Why does this make sense?

- Intuitively, “maximum margin” gives greatest robustness to errors in the data
  - Generalization error is inversely proportional to margin (Bartlett and Shawe-Taylor 1998)
- The classifier depends on only a few data points, so it is
  - “Sparse” (has few parameters to learn)
  - Efficient to evaluate

Calculating the Margin



- Plus-plane = $\mathbf{w}\cdot\mathbf{x} + b = +1$
- Minus-plane = $\mathbf{w}\cdot\mathbf{x} + b = -1$

Classify as..	$+1$	if	$\mathbf{w}\cdot\mathbf{x} + b \geq 1$
	-1	if	$\mathbf{w}\cdot\mathbf{x} + b \leq -1$

Calculating the Margin

- First note that $\mathbf{w}$ is perpendicular to the plane $\mathbf{w}\cdot\mathbf{x}+b=0$
- Why?
  - Pick $\mathbf{u}$, $\mathbf{v}$ on plane
  - $\mathbf{w}\cdot(\mathbf{u}-\mathbf{v})=\mathbf{w}\cdot\mathbf{u} - \mathbf{w}\cdot\mathbf{v}=(-b) -(-b)=0$
- So $\mathbf{w}$ is also perpendicular to the plus and minus planes

Calculating the Margin

- Choose an arbitrary point $\mathbf{x}^{+}$ on the plus plane and its nearest point $\mathbf{x}^{-}$ on the minus plane
- Notice that $\mathbf{x}^{+} - \mathbf{x}^{-} = \lambda\mathbf{w}$ and so $M = \| \lambda\mathbf{w} \|_2$

$$\mathbf{w} \bullet \mathbf{x}^{+} + b = 1$$

$$\mathbf{w} \bullet (\mathbf{x}^{-} + \lambda\mathbf{w}) + b = 1$$

$$\lambda\mathbf{w} \bullet \mathbf{w} = 2; \lambda = \frac{2}{\mathbf{w} \bullet \mathbf{w}} = \frac{2}{\|\mathbf{w}\|^2}$$

$$M = \|\lambda\mathbf{w}\| = \frac{2}{\|\mathbf{w}\|}$$

So maximizing the margin is equivalent here to minimizing the norm of the parameter vector! Also a rationale behind other overfitting control methods like weight decay

Problem Formulation

- On the training set,
  - Maximize the margin
  - *While respecting the labels of the training examples*
- Maximize the margin

$$\max_{\mathbf{w},b} \frac{2}{\|\mathbf{w}\|} = \min_{\mathbf{w},b} \frac{\|\mathbf{w}\|^2}{2}$$

Problem Formulation

- While respecting the labels of training examples---these are *constraints*

$$\mathbf{w}\bullet\mathbf{x}_i + b \geq 1 \text{ if } y_i = 1$$
$$\mathbf{w}\bullet\mathbf{x}_i + b \leq -1 \text{ if } y_i = -1$$
$$\Rightarrow$$
$$y_i(\mathbf{w}\bullet\mathbf{x}_i + b) \geq 1$$

One such constraint for each example

Problem Formulation

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{so that } \forall i, y_i(\mathbf{w} \bullet \mathbf{x}_i + b) \geq 1$$

- Called a "quadratic program"
  - Many methods to solve, e.g. successive linearization
- Has globally unique solution! (convexity)
- So are we done?

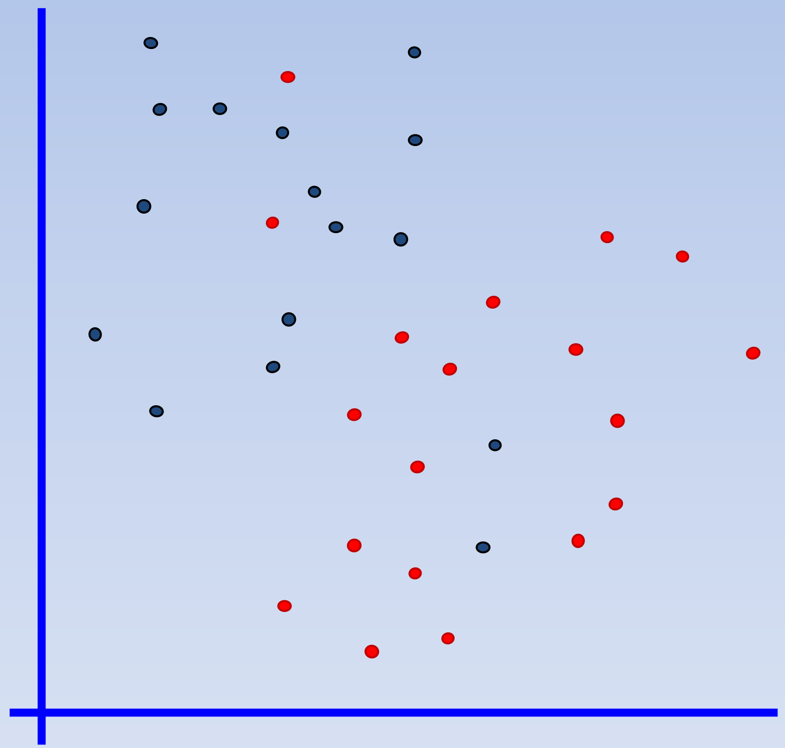
Problem Formulation

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{so that } \forall i, y_i(\mathbf{w}\bullet\mathbf{x}_i + b) \geq 1$$

- Called a "quadratic program"
  - Many methods to solve, e.g. successive linearization
- Has globally unique solution! (convexity)
- So are we done?

Linearly Inseparable Data



What happens to the QP in this case?

Linearly Inseparable Data

- Normally, we have:

$$y_i(\mathbf{w}\bullet\mathbf{x}_i + b) \geq 1$$

- So to allow for a misclassified point:

$$y_i(\mathbf{w}\bullet\mathbf{x}_i + b) \leq 1$$

$$\text{or } y_i(\mathbf{w}\bullet\mathbf{x}_i + b) + \xi_i \geq 1, \xi_i \geq 0$$

Free “slack” variables. The optimizer will find these values as well.

Problem Formulation, LI Data

$$\min_{\mathbf{w},b,\xi_i} \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{so that } \forall i, y_i(\mathbf{w} \bullet \mathbf{x}_i + b) + \xi_i \geq 1, \xi_i \geq 0$$

Oops, doesn't work!! Try $\mathbf{w}$=0, b=0, ξ_i=1. Can't just allow for misclassified points---must *minimize the number of misclassified points as well*!

LI Data, Attempt 2

- Want:

$$\min_{\mathbf{w},b,\xi_i} \frac{1}{2}\|\mathbf{w}\|^2 + [Number\_of\_errors]$$

$$\text{so that } \forall i,\, y_i(\mathbf{w}\bullet\mathbf{x}_i + b) + \xi_i \geq 1,\, \xi_i \geq 0$$

- But this is problematic, because number of errors is not a differentiable quantity

LI Data, Attempt 2

- We know that:

$$y_i(\mathbf{w} \bullet \mathbf{x}_i + b) + \xi_i \geq 1, \xi_i \geq 0,$$

$$\text{So } \xi_i \geq 1 - y_i(\mathbf{w} \bullet \mathbf{x}_i + b)$$

- So $0 \leq \xi_i < 1$ for correctly classified points, and $\xi_i \geq 1$ for incorrectly classified points
- So $\sum \xi_i$ is an upper bound on the number of errors
  - This is a differentiable quantity we can minimize

Final Formulation

Tradeoff between generalization and error

Slack for i^{th} example

$$\min_{\mathbf{w},b,\xi_i} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i$$

$$\text{so that } \forall i,\, y_i(\mathbf{w}\bullet\mathbf{x}_i + b) + \xi_i \geq 1$$

$$\text{and} \qquad \forall i,\, \xi_i \geq 0$$

Nonlinear (in $\mathbf{x}$) Classifiers

- So far, we have looked at SVMs linear in $\mathbf{x}$
- How do we learn decision surfaces nonlinear in $\mathbf{x}$?

SVM Formulation

$$\min_{\mathbf{w},b,\xi_i} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i$$

$$\text{so that } \forall i, y_i(\mathbf{w} \bullet \varphi(\mathbf{x}_i) + b) + \xi_i \geq 1$$

$$\text{and} \qquad \forall i, \xi_i \geq 0$$

Nonlinear (in $\mathbf{x}$) Classifiers

- But it turns out we need not explicitly compute $\boldsymbol{\varphi}(x)$ at all!
  - Using "kernels"
  - "Implicit feature map"
  - Computational savings

- To get this, we will build the *dual form* of the linear SVM's QP using the "Generalized Lagrangian"

Recall: Duality in Linear Programming

- From any "primal" LP, we can derive a "dual" LP in the following way:

$$\min_{\mathbf{x}} c^T \mathbf{x}$$
$$s.t.\ \ A\mathbf{x} \geq b$$
$$\mathbf{x} \geq 0$$

"Primal" problem

$$\max_{\mathbf{u}} b^T \mathbf{u}$$
$$s.t.\ \ A^T \mathbf{u} \leq c$$
$$\mathbf{u} \geq 0$$

"Dual" problem

Generalized Lagrangian

- Consider the following problem:

$$\min_{w} f(w)$$
$$\text{so that } g_i(w) \leq 0$$
$$\text{and} \qquad h_j(w) = 0$$

- The generalized Lagrangian is defined by:

$$\ell(w, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(w) + \sum_i \alpha_i g_i(w) + \sum_j \beta_j h_j(w)$$

"Langrangian multipliers"

For the linearly-separable SVM

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{so that } \forall i, -\left[y_i(\mathbf{w}\bullet\mathbf{x}_i + b) - 1\right] \leq 0$$

$$\therefore \ell(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_i \alpha_i \left[y_i(\mathbf{w}\bullet\mathbf{x}_i + b) - 1\right]$$

Lagrange Duality

- Consider $P(w) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\boldsymbol{\alpha}\geq 0} \ell(w, \boldsymbol{\alpha}, \boldsymbol{\beta})$

$$P(w) = \max_{\alpha,\beta:\alpha\geq 0} f(w) + \sum_i \alpha_i g_i(w) + \sum_j \beta_j h_j(w)$$

"Primal" problem

$$= \begin{cases} f(w) \text{ if constraints on } g \text{ and } h \text{ are met} \\ \infty \text{ else} \end{cases}$$

- So the original problem can be written as

$$\min_w P(w) = \min_w \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\boldsymbol{\alpha}\geq 0} \ell(w, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

Lagrange Duality

- Consider $\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\boldsymbol{\alpha}\geq 0} D(\boldsymbol{\alpha},\boldsymbol{\beta}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\boldsymbol{\alpha}\geq 0} \min_{w} \ell(w,\boldsymbol{\alpha},\boldsymbol{\beta})$
- This is the dual formulation corresponding to P
  - So starting with ℓ, we can *derive* the dual for a primal problem

For the linearly-separable SVM

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{so that } \forall i, -\left[y_i(\mathbf{w}\bullet\mathbf{x}_i + b) - 1\right] \leq 0$$

$$\therefore \ell(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_i \alpha_i \left[y_i(\mathbf{w}\bullet\mathbf{x}_i + b) - 1\right]$$

Linearly-separable SVM, Dual Form

$$\ell(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_i \alpha_i \left[y_i (\mathbf{w} \bullet \mathbf{x}_i + b) - 1 \right]$$

$$\nabla_{\mathbf{w}} \ell(\mathbf{w}, b, \boldsymbol{\alpha}) = \mathbf{w} - \sum_i \alpha_i y_i \mathbf{x}_i = 0$$

$$\therefore \mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$$

$$\nabla_b \ell(\mathbf{w}, b, \boldsymbol{\alpha}) = \sum_i \alpha_i y_i = 0$$

Substitute for $\mathbf{w}$ in ℓ

Linearly-separable SVM, Dual Form

$$\ell(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_i \alpha_i \left[y_i(\mathbf{w} \bullet \mathbf{x}_i + b) - 1 \right]$$

$$\mathbf{w} = \sum_j \alpha_j y_j \mathbf{x}_j ; \sum_i \alpha_i y_i = 0$$

Substitute for $\mathbf{w}$ in ℓ

$$D(\boldsymbol{\alpha}) = \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j \mathbf{x}_i \bullet \mathbf{x}_j + \sum_i \alpha_i - \sum_{i,j} y_i y_j \alpha_i \alpha_j \mathbf{x}_i \bullet \mathbf{x}_j - b \sum_i \alpha_i y_i$$

$$= \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j \mathbf{x}_i \bullet \mathbf{x}_j$$

Linearly-separable SVM, Dual Form

$$\max_{\alpha} D(\boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j \mathbf{x}_i \bullet \mathbf{x}_j$$

$$\text{so that } \boldsymbol{\alpha} \geq 0, \sum_i \alpha_i y_i = 0$$

From derivative w.r.t b

Karush-Kuhn-Tucker conditions

- At the optimal primal/dual solution, the following conditions will hold:

$$\nabla_{\mathbf{w},b}\ell(\mathbf{w}^*, b^*, \boldsymbol{\alpha}^*) = 0$$

Gradient at solution is zero

$$-\left[y_i(\mathbf{w}^* \bullet \mathbf{x}_i + b^*) - 1\right] \leq 0$$

$$\alpha_i^* \geq 0$$

All constraints satisfied

$$\alpha_i^*\left[y_i(\mathbf{w}^* \bullet \mathbf{x}_i + b^*) - 1\right] = 0$$

KKT dual complementarity
If i^{th} LM is positive, the i^{th} constraint is "active", i.e. zero
These are the **support vectors**

These conditions are **necessary and sufficient!**