Foundations of Machine Learning

Module 5:

Part B: Introduction to Support Vector Machine

Sudeshna Sarkar
IIT Kharagpur

Support Vector Machines

- SVMs have a clever way to prevent overfitting
- They can use many features without requiring too much computation.

Logistic Regression and Confidence

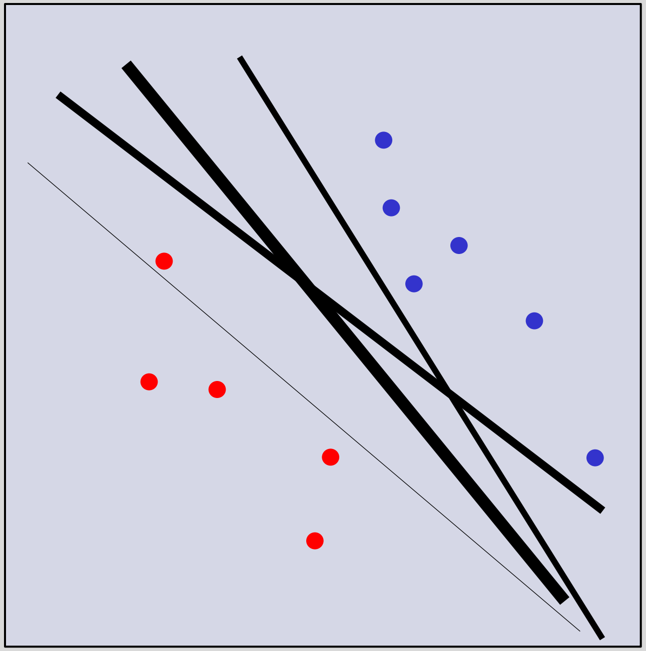
- Logistic Regression:

$p(y=1|x) = h_\beta(x) = g(\beta^T x)$

- Predict 1 on an input x iff $h_\beta(x) \geq 0.5$, equivalently, $\beta^T x \geq 0$
- The larger the value of $h_\beta(x)$, the larger is the probability, and higher the confidence.
- Similarly, confident prediction of $y = 0$ if $\beta^T x \ll 0$
- More confident of prediction from points (instances) located far from the decision surface.

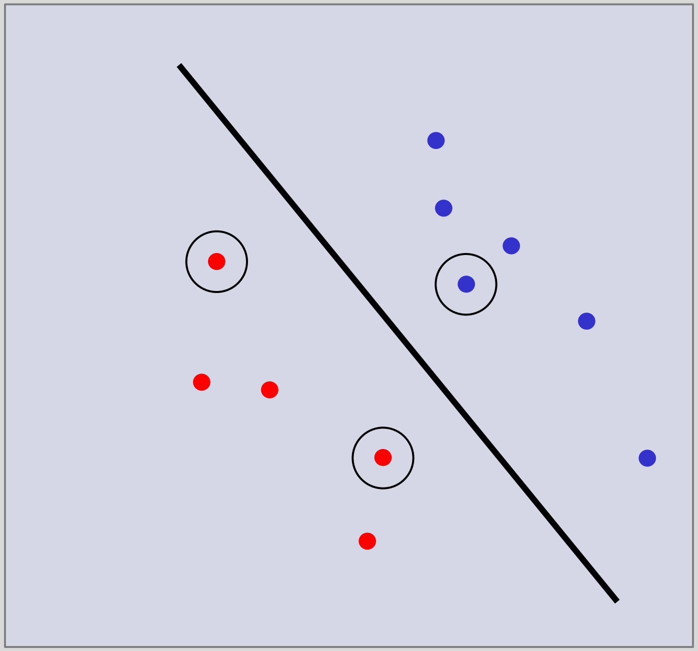
Preventing overfitting with many features

- Suppose a big set of features.
- What is the best separating line to use?
- Bayesian answer:
  - Use all
  - Weight each line by its posterior probability
- Can we approximate the correct answer efficiently?



Support Vectors

- The line that maximizes the minimum margin.
- This maximum-margin separator is determined by a subset of the datapoints.
  - called “support vectors”.
  - we use the support vectors to decide which side of the separator a test case is on.



The support vectors are indicated by the circles around them.

Functional Margin

- Functional Margin of a point (x_i,y_i) wrt (w,b)
  - Measured by the distance of a point (x_i,y_i) from the decision boundary (w,b)

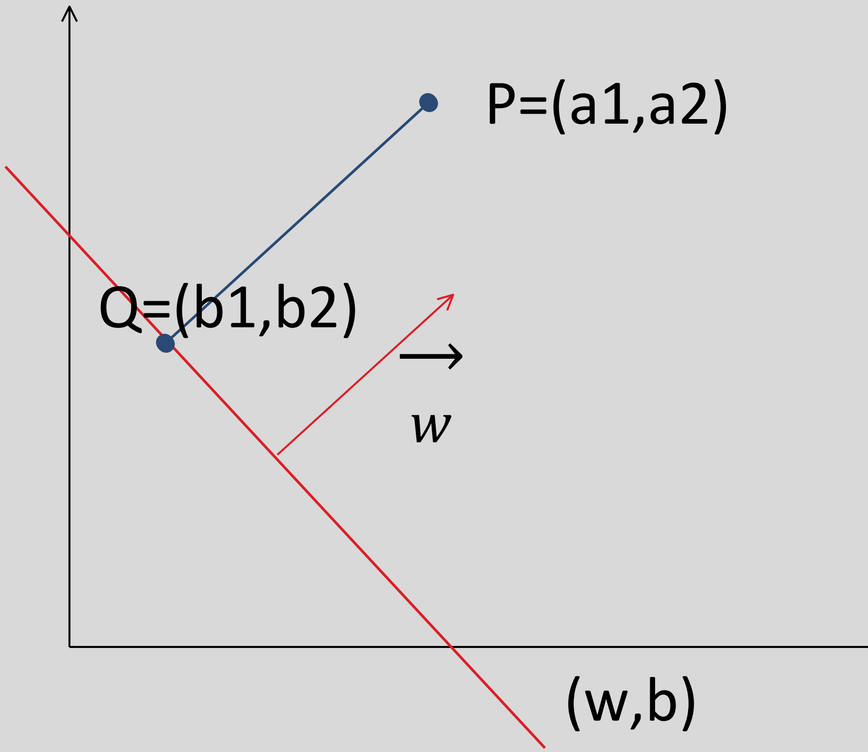
  $$\gamma^i = y_i(w^T x_i + b)$$

  - Larger functional margin →more confidence for correct prediction
  - Problem: w and b can be scaled to make this value larger
- Functional Margin of training set $\{(x_1,y_1), (x_2,y_2), \ldots, (x_m,y_m)\}$ wrt (w,b) is

$$\gamma = \min_{1 \le i \le m} \gamma^i$$

Geometric Margin

- For a decision surface (w,b)
- the vector orthogonal to it is given by w.
- The unit length orthogonal vector is $\frac{w}{||w||}$
- $P = Q + \gamma \frac{w}{||w||}$



Geometric Margin

$$P = Q + \gamma \frac{w}{\|w\|}$$

$$(b1,b2) = (a1,a2) - \gamma \frac{w}{\|w\|}$$

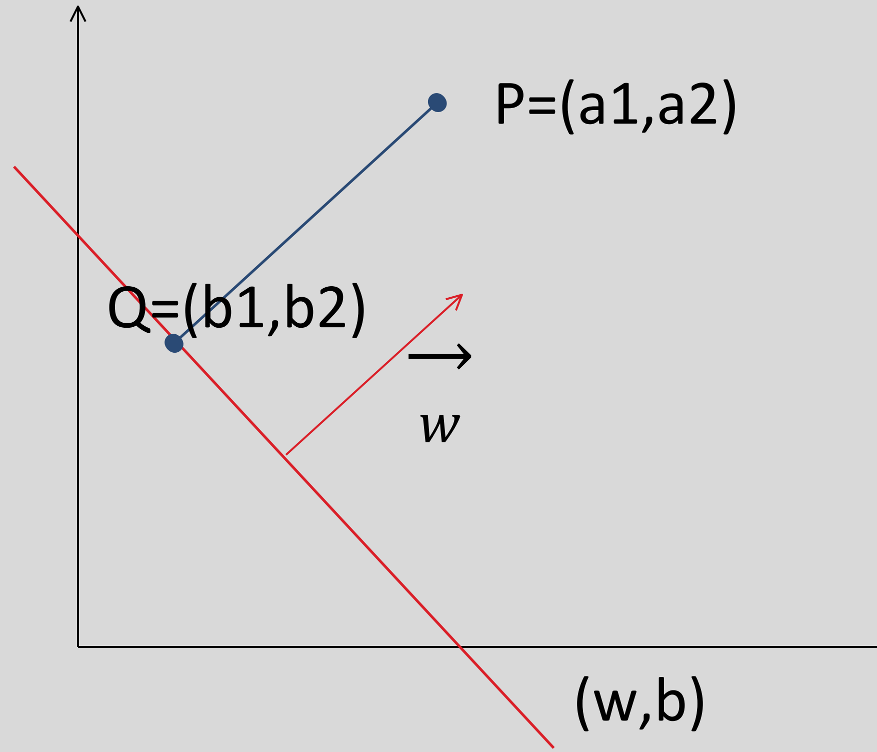
$$\rightarrow\ w^T\left((a1,a2) - \gamma \frac{w}{\|w\|}\right) + b = 0$$

$$\rightarrow \gamma = \frac{w^T(a1,a2)+b}{\|w\|}$$

$$= \frac{w}{\|w\|}^T (a1,a2) + \frac{b}{\|w\|}$$

$$= \frac{w}{\|w\|}^T (a1,a2) + \frac{b}{\|w\|}$$

$$\gamma = y.(\frac{w}{\|w\|}^T (a1,a2) + \frac{b}{\|w\|})$$



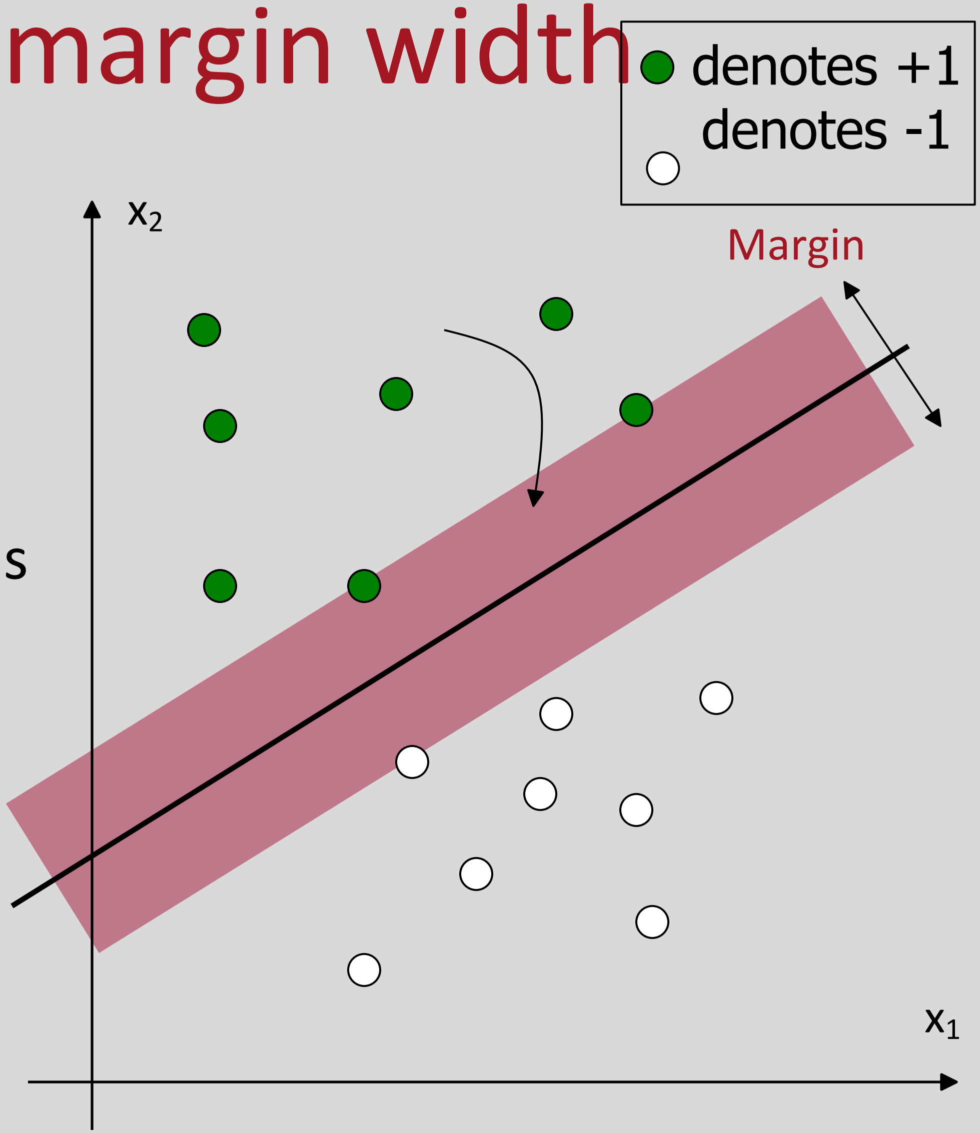
Geometric margin : $\|w\| = 1$

Geometric margin of (w,b) wrt S={ (x_1,y_1), (x_2,y_2), ..., (x_m,y_m) }

-- smallest of the geometric margins of individual points.

Maximize margin width

- Assume linearly separable training examples.
- The classifier with the maximum margin width is robust to outliners and thus has strong generalization ability



Maximize Margin Width

- Maximize $\frac{\gamma}{||w||}$ subject to
- $y_i\left(w^T x_i + b\right) \geq \gamma$ for $i = 1,2, .. ,m$
- Scale so that $\gamma = 1$
- Maximizing $\frac{1}{||w||}$ is the same as minimizing $||w||^2$
- Minimize $\boldsymbol{w}.\boldsymbol{w}$ subject to the constraints
- for all (x_i, y_i), $i = 1,\ldots,m$:

  $w^T x_i + b \geq 1$ if $y_i = 1$

  $w^T x_i + b \leq -1$ if $y_i = -1$

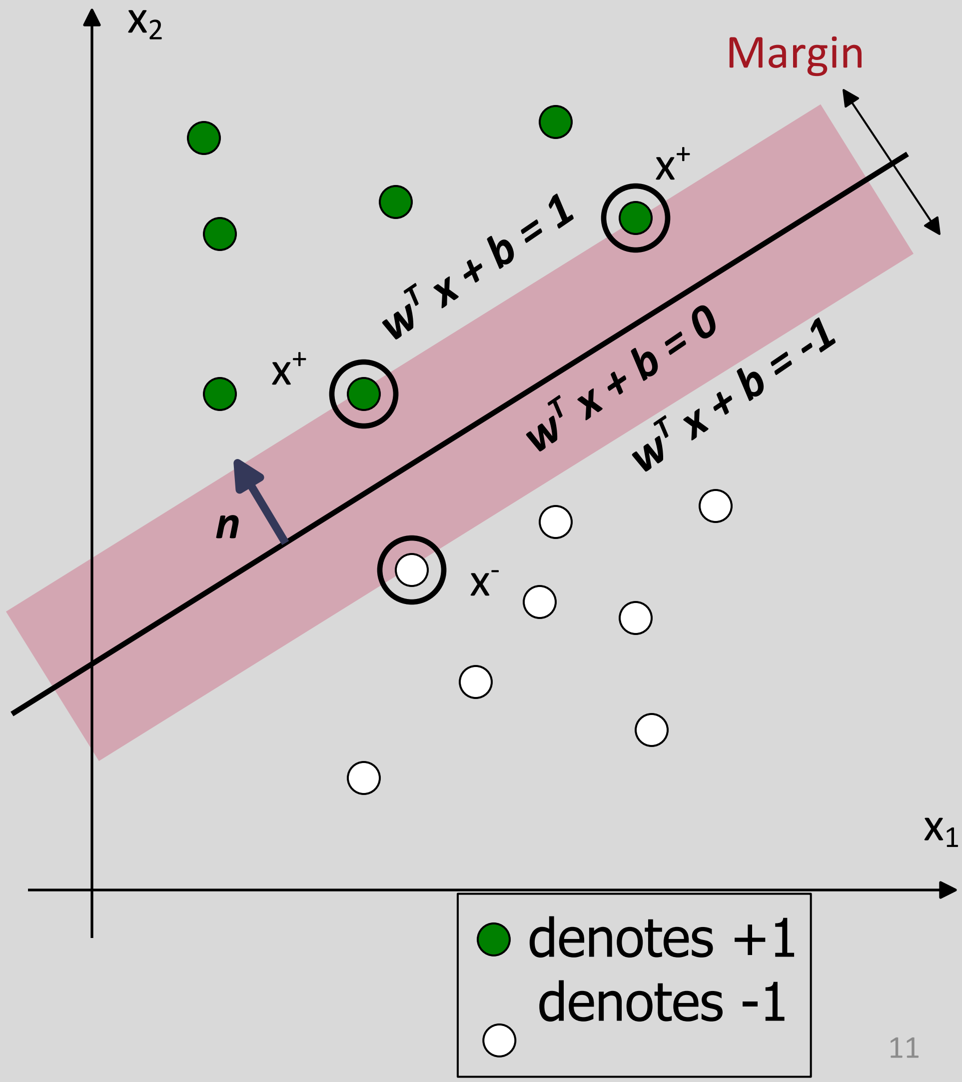
Large Margin Linear Classifier

- Formulation:

$$\text{minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2$$

such that

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



Solving the Optimization Problem

$$\text{minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$$

- Optimization problem with convex quadratic objectives and linear constraints
- Can be solved using QP.
- Lagrange duality to get the optimization problem's dual form,
  - Allow us to use kernels to get optimal margin classifiers to work efficiently in very high dimensional spaces.
  - Allow us to derive an efficient algorithm for solving the above optimization problem that will typically do much better than generic QP software.