Machine learning: lecture 7

Tommi S. Jaakkola
MIT CSAIL
tommi@csail.mit.edu

Topics

- Logistic regression
  - conditional family, quantization
  - regularization
  - penalized log-likelihood
- Non-probabilistic classification: support vector machine
  - linear discrimination
  - regularization and "optimal" hyperplane
  - optimization via Lagrange multipliers

Review: logistic regression

- Consider a simple logistic regression model

$$P(y = 1|x, \mathbf{w}) = g(w_0 + w_1 x)$$

parameterized by $\mathbf{w} = (w_0, w_1)$. We assume that $x \in [-1, 1]$ (or more generally that the input remains bounded).

Review: logistic regression

- Consider a simple logistic regression model

$$P(y = 1|x, \mathbf{w}) = g(w_0 + w_1 x)$$

parameterized by $\mathbf{w} = (w_0, w_1)$. We assume that $x \in [-1, 1]$ (or more generally that the input remains bounded).

- We view this model as a set of possible conditional distributions (family of conditionals):

$$P(y = 1|x, \mathbf{w}) \;\; = \;\; g(w_0 + w_1 x), \;\; \mathbf{w} = [w_0, w_1]^T \in \mathcal{R}^2$$

Review: logistic regression

- Consider a simple logistic regression model

$$P(y = 1|x, \mathbf{w}) = g(w_0 + w_1 x)$$

parameterized by $\mathbf{w} = (w_0, w_1)$. We assume that $x \in [-1, 1]$ (or more generally that the input remains bounded).

- We view this model as a set of possible conditional distributions (family of conditionals):

$$P(y = 1|x, \mathbf{w}) \;\;=\;\; g(w_0 + w_1 x),\;\; \mathbf{w} = [w_0, w_1]^T \in \mathcal{R}^2$$

- It does not matter how the conditionals are parameterized. For example, the following definition gives rise to the same family:

$$P(y = 1|x, \tilde{\mathbf{w}}) \;\;=\;\; g\big(\tilde{w}_0 + (\tilde{w}_2 - \tilde{w}_1)x\big),\;\; \tilde{\mathbf{w}} = [\tilde{w}_0, \tilde{w}_1, \tilde{w}_2]^T \in \mathcal{R}^3$$

Review: "choices" in logistic regression

- We are interested in "quantizing" the set of conditionals

$$P(y = 1|x, \mathbf{w}) \;\; = \;\; g(w_0 + w_1 x), \;\mathbf{w} = [w_0, w_1]^T \in \mathcal{R}^2$$

by finding a discrete representative set that essentially captures all the possible conditional distributions we have in this family.

Review: "choices" in logistic regression

- We are interested in "quantizing" the set of conditionals

$$P(y = 1|x, \mathbf{w}) \;\; = \;\; g(w_0 + w_1 x), \;\; \mathbf{w} = [w_0, w_1]^T \in \mathcal{R}^2$$

  by finding a discrete representative set that essentially captures all the possible conditional distributions we have in this family.

- We can represent this discrete set in terms of different parameter choices $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_\infty$

Review: "choices" in logistic regression

- We are interested in "quantizing" the set of conditionals

$$P(y=1|x,\mathbf{w}) \;\; = \;\; g(w_0+w_1 x), \;\; \mathbf{w} = [w_0, w_1]^T \in \mathcal{R}^2$$

by finding a discrete representative set that essentially captures all the possible conditional distributions we have in this family.

- We can represent this discrete set in terms of different parameter choices $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_\infty$

- Any conditional $P(y|x,\mathbf{w})$ should be close to one of the discrete choices $P(y|x,\mathbf{w}_j)$ in the sense that they make "similar" predictions for all inputs $x \in [-1, 1]$:

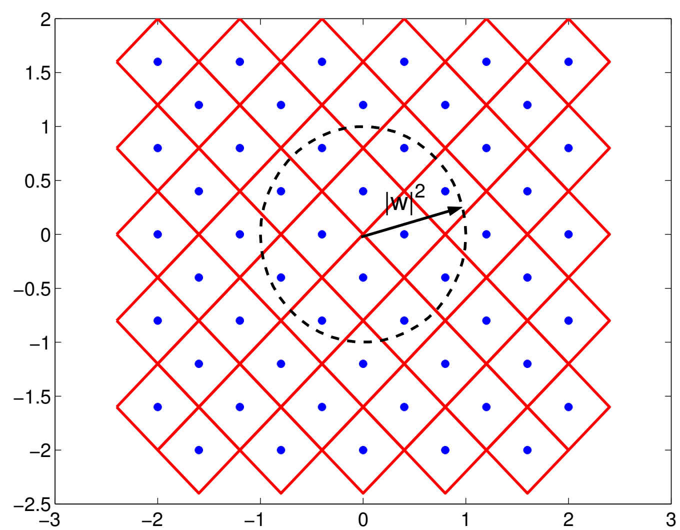
$$|\log P(y=1|x,\mathbf{w}) - \log P(y=1|x,\mathbf{w}_j)| \leq \epsilon$$

Review: "choices" in logistic regression

- We can view the discrete parameter choices $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_\infty$ as "centroids" of regions in the parameter space such that within each region

$$|\log P(y=1|x, \mathbf{w}) - \log P(y=1|x, \mathbf{w}_j)| \leq \epsilon$$

for all $x \in [-1, 1]$

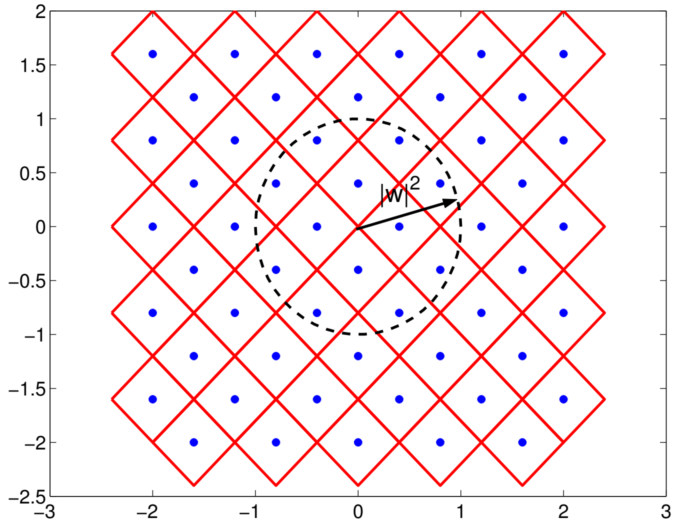


Review: "choices" in logistic regression

- We can view the discrete parameter choices $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_\infty$ as "centroids" of regions in the parameter space such that within each region

$$|\log P(y=1|x,\mathbf{w}) - \log P(y=1|x,\mathbf{w}_j)| \leq \epsilon$$

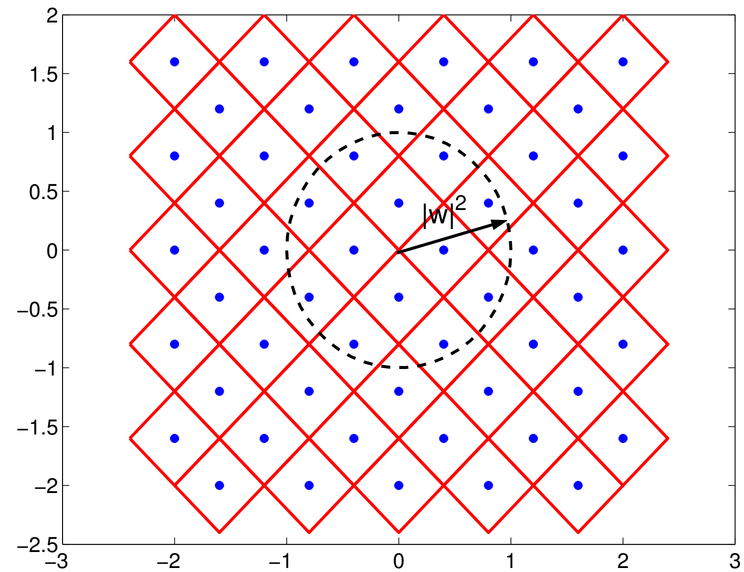
for all $x \in [-1, 1]$



- Regularization means limiting the number of choices we have in this family. For example, we can constrain $\|\mathbf{w}\| \leq C$.

Regularized logistic regression

- We can regularize the models by imposing a penalty in the estimation criterion that encourages $\|\mathbf{w}\|$ to remain small.

  Maximum penalized log-likelihood criterion:

$$l(D; \mathbf{w}, \lambda) = \sum_{i=1}^{n} \log P(y_i | \mathbf{x}_i, \mathbf{w}) - \frac{\lambda}{2} \|\mathbf{w}\|^2$$

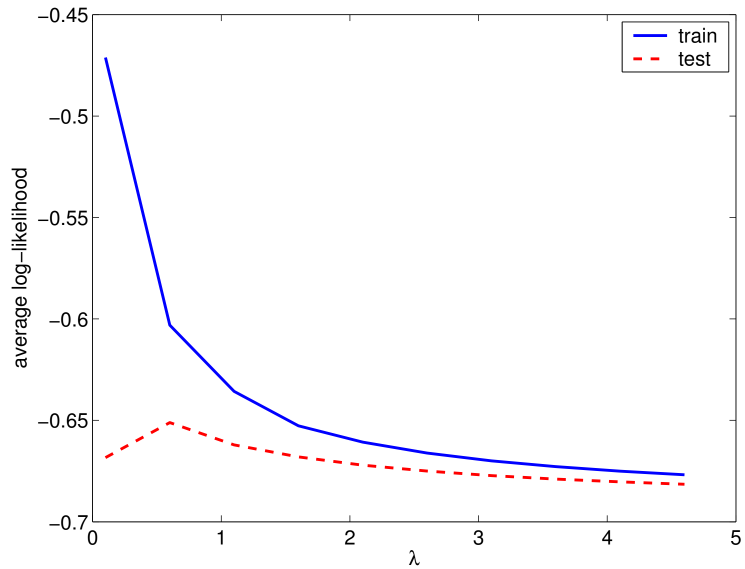
  where larger values of λ impose stronger regularization.

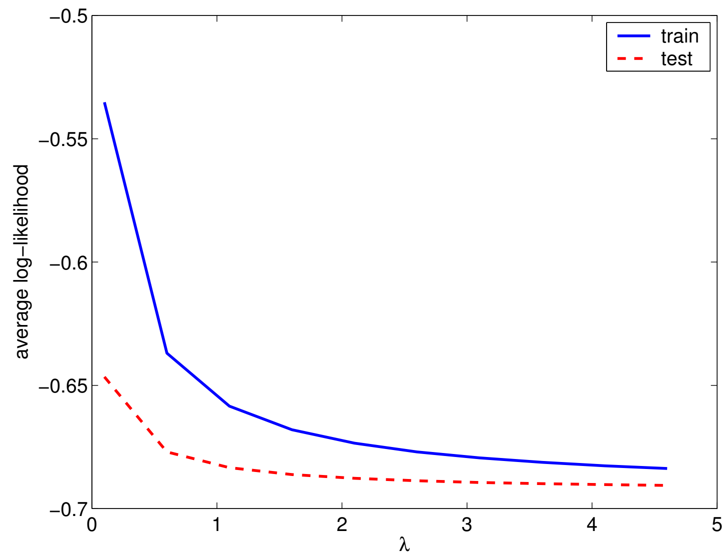
- More generally, we can assign penalties based on prior distributions over the parameters, i.e., add $\log P(\mathbf{w})$ in the log-likelihood criterion.

Regularized logistic regression

- How do the training/test conditional log-likelihoods behave as a function of the regularization parameter λ?

$$l(D; \mathbf{w}, \lambda) = \sum_{i=1}^{n} \log P(y_i | \mathbf{x}_i, \mathbf{w}) - \frac{\lambda}{2} \|\mathbf{w}\|^2$$





Topics

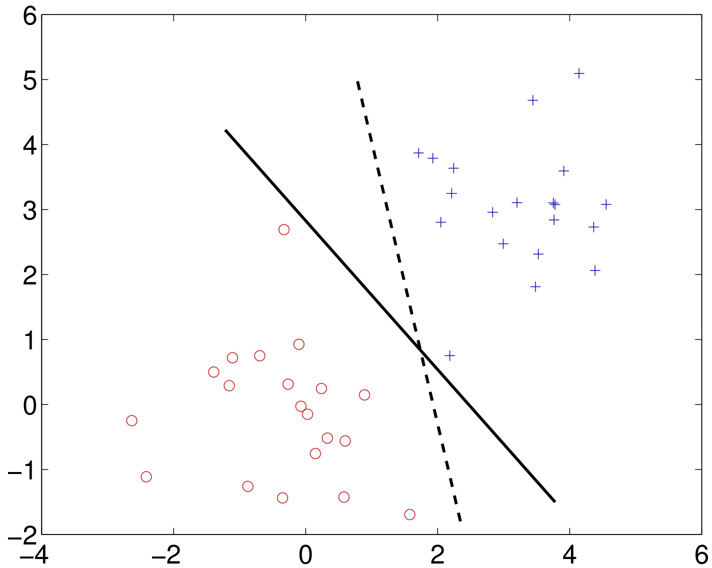
- Logistic regression
  - conditional family, quantization
  - regularization
  - penalized log-likelihood
- Non-probabilistic classification: support vector machine
  - linear discrimination
  - regularization and “optimal” hyperplane
  - optimization via Lagrange multipliers

Non-probabilistic classification

- Consider a binary classification task with $y = \pm 1$ labels (not $0/1$ as before) and linear *discriminant* functions:

$$f(\mathbf{x}; w_0, \mathbf{w}) = w_0 + \mathbf{w}^T \mathbf{x}$$

parameterized by $\{w_0, \mathbf{w}\}$. The label we predict for each example is given by the sign of the linear function $w_0 + \mathbf{w}^T \mathbf{x}$.

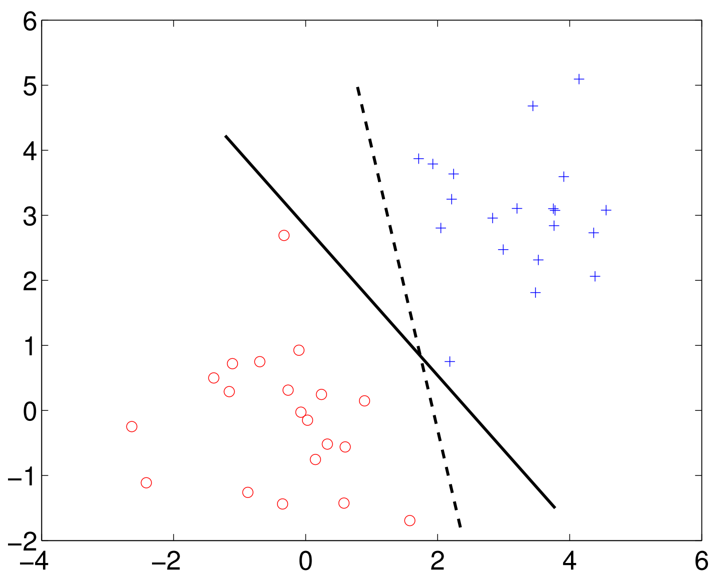


Linear classification

- When training examples are *linearly separable* we can set the parameters of a linear classifier so that all the training examples are classified correctly:

$$y_i \left[w_0 + \mathbf{w}^T \mathbf{x}_i \right] > 0, \; i = 1, \ldots, n$$

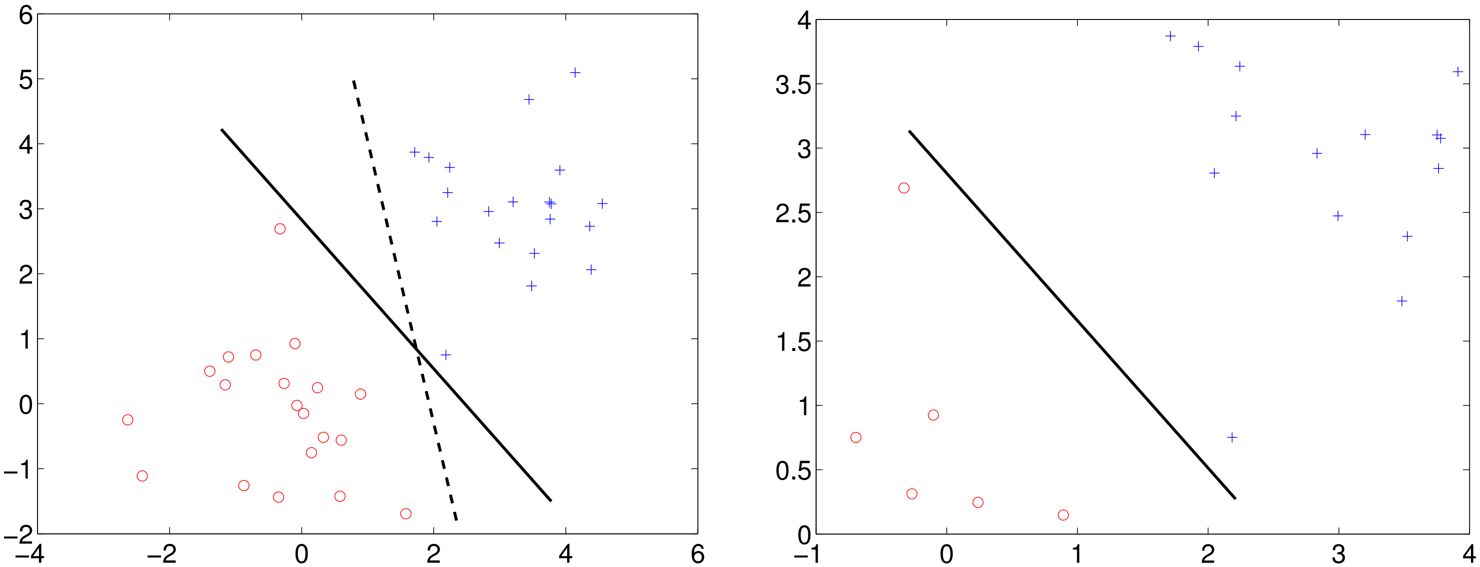
(the sign of the label agrees with the sign of the linear function $w_0 + \mathbf{w}^T \mathbf{x}$)



Classification and margin

- We can try to find a unique solution by requiring that the training examples are classified correctly with a non-zero "margin"

$$y_i \left[w_0 + \mathbf{w}^T \mathbf{x}_i \right] - 1 \geq 0, \; i = 1, \ldots, n$$

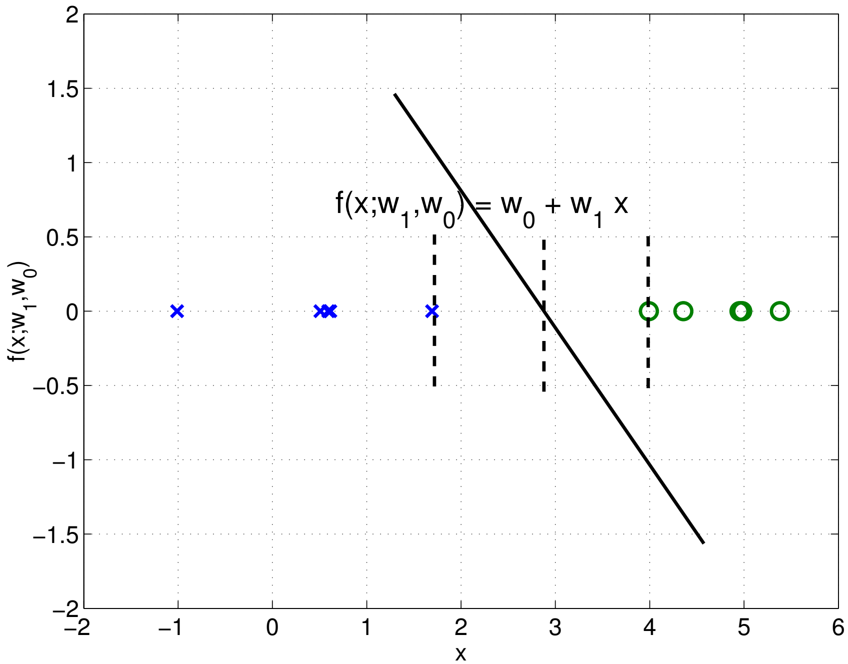


The margin should be defined in terms of the distance from the boundary to the examples rather than based on the value of the linear function.

Margin and slope

- One dimensional example: $f(x; w_1, w_0) = w_0 + w_1 x$.
  Relevant constraints:

$$
\begin{aligned}
1\left[w_0 + w_1 x^+\right] - 1 &\geq 0 \\
-1\left[w_0 + w_1 x^-\right] - 1 &\geq 0
\end{aligned}
$$

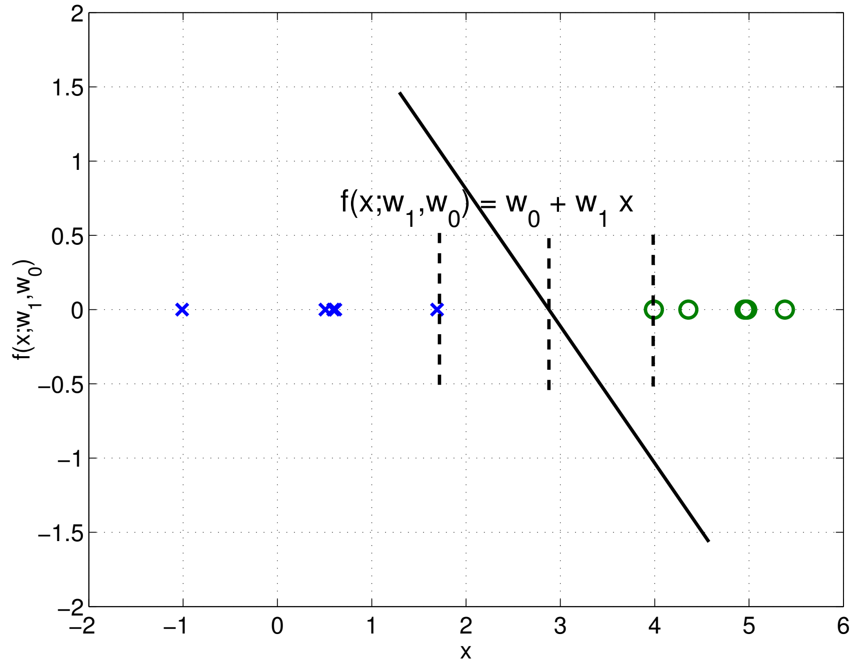


Margin and slope

- One dimensional example: $f(x; w_1, w_0) = w_0 + w_1 x$. Relevant constraints:

$$1\left[w_0 + w_1 x^{+}\right] - 1 \geq 0$$
$$-1\left[w_0 + w_1 x^{-}\right] - 1 \geq 0$$

We obtain the maximum separation at the mid point with margin $|x^{+} - x^{-}|/2$.

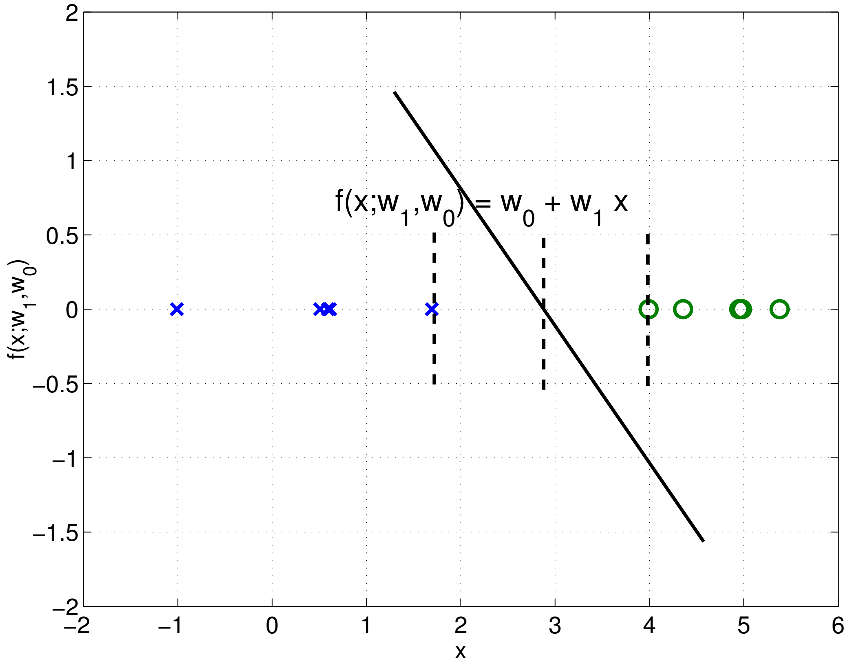


Margin and slope

- One dimensional example: $f(x; w_1, w_0) = w_0 + w_1 x$. Relevant constraints:

$$1\,[w_0 + w_1 x^+] - 1 \;\geq\; 0$$
$$-1\,[w_0 + w_1 x^-] - 1 \;\geq\; 0$$

  We obtain the maximum separation at the mid point with margin $|x^+ - x^-|/2$.



- This is the only possible solution if we minimize the slope $|w_1|$ subject to the constraints. At the optimum

$$|w_1^*| = \frac{1}{|x^+ - x^-|/2} = \frac{1}{\text{margin}}$$

Support vector machine

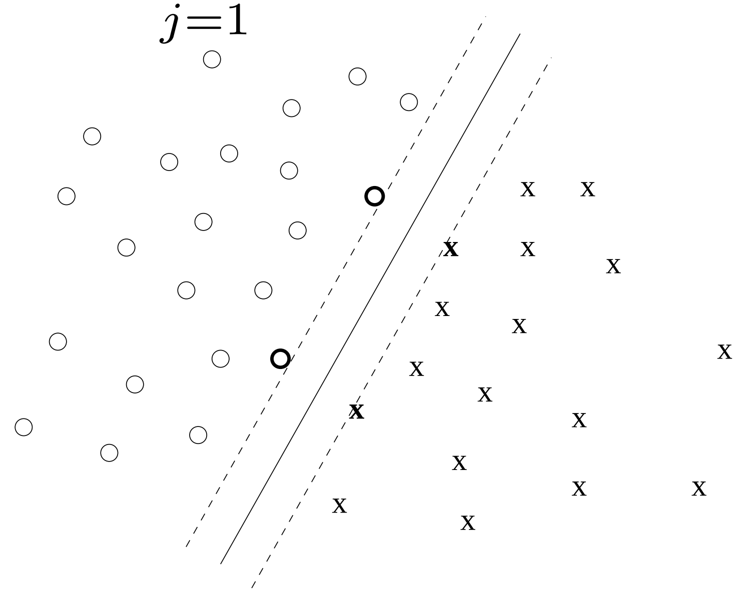
- We minimize a regularization penalty

$$\|\mathbf{w}\|^2/2 = \mathbf{w}^T\mathbf{w}/2 = \sum_{j=1}^{d} w_i^2/2$$

subject to the classification constraints

$y_i\left[w_0 + \mathbf{w}^T\mathbf{x}_i\right] - 1 \geq 0,$

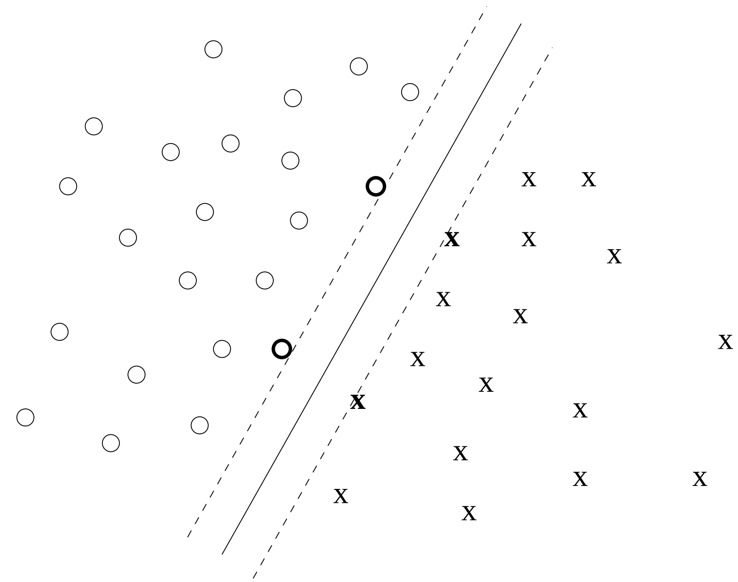
for $i = 1, \ldots, n$.



- Analogously to the one dimensional case, the "slope" is again related to the margin: $\|\mathbf{w}^*\| = 1/\text{margin}$.

Support vector machine cont'd

- Only a few of the classification constraints are relevant



- We could in principle define the solution on the basis of only a small subset of the training examples called "support vectors"

Support vector machine: solution

- We find the optimal setting of $\{w_0, \mathbf{w}\}$ by introducing *Lagrange multipliers* $\alpha_i \geq 0$ for the inequality constraints

- We *minimize*

$$J(\mathbf{w}, w_0, \alpha) = \|\mathbf{w}\|^2/2 - \sum_{i=1}^{n} \alpha_i \left(y_i \left[w_0 + \mathbf{w}^T \mathbf{x}_i \right] - 1 \right)$$

with respect to $\mathbf{w}, w_0$. $\{\alpha_i\}$ ensure that the classification constraints are indeed satisfied.

For fixed $\{\alpha_i\}$

$$\begin{aligned} \frac{\partial}{\partial \mathbf{w}} J(\mathbf{w}, w_0, \alpha) &= \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0 \\ \frac{\partial}{\partial w_0} J(\mathbf{w}, w_0, \alpha) &= -\sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned}$$

Solution

- Substituting the solution $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$ back into the objective leaves us with the following (dual) optimization problem over the Lagrange multipliers:

  We *maximize*

  $$J(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j)$$

  subject to the constraints

  $$\alpha_i \geq 0, \quad i = 1, \ldots, n, \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

  (For non-separable problems we have to limit $\alpha_i \leq C$)

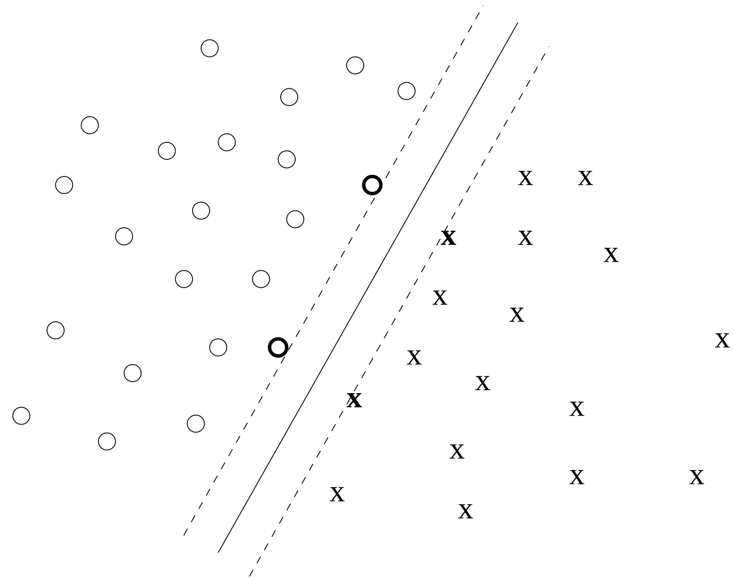
- This is a *quadratic programming problem*

Support vector machines

- Once we have the Lagrange multipliers $\{\hat{\alpha}_i\}$, we can reconstruct the parameter vector $\hat{\mathbf{w}}$ as a weighted combination of the training examples:

$$\hat{\mathbf{w}} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \mathbf{x}_i$$

  where the "weight" $\hat{\alpha}_i = 0$ for all but the *support vectors* (SV)



- The decision boundary has an interpretable form

$$\hat{\mathbf{w}}^T\mathbf{x} + \hat{w}_0 = \sum_{i \in SV} \hat{\alpha}_i \, y_i \, (\mathbf{x}_i^T \mathbf{x}) + \hat{w}_0 = f(\mathbf{x}; \hat{\alpha}, \hat{w}_0)$$

Interpretation of support vector machines

- To use support vector machines we have to specify only the inner products (or *kernel*) between the examples $(\mathbf{x}_i^T \mathbf{x})$
- The weights $\{\alpha_i\}$ associated with the training examples are solved by enforcing the classification constraints.

  $\Rightarrow$ sparse solution
- We make decisions by comparing each new example $\mathbf{x}$ with **only** the support vectors $\{\mathbf{x}_i\}_{i \in SV}$:

$$\hat{y} = \mathrm{sign}\left(\sum_{i \in SV} \hat{\alpha}_i \, y_i \, (\mathbf{x}_i^T \mathbf{x}) + \hat{w}_0 \right)$$