

CS195-5 : Introduction to Machine Learning

Lecture 14

Greg Shakhnarovich

October 6, 2006

Announcements

- 10/9: no class (Columbus Day)
- 10/13: Guest lecture: Meinolf Sellman
 - Optimization and Lagrange multipliers
- 10/16: no class.
- 10/18: Guest lecture: Chad Jenkins
 - Robot learning, intro to unsupervised and reinforcement learning.

Review

- Regularization for model parametrized by \mathbf{w} , trained on data D :

$$\hat{\mathbf{w}} = \operatorname{argmax}_{\mathbf{w}} \log\text{-likelihood}(D; \mathbf{w}) - \lambda \cdot \text{penalty}(\mathbf{w}).$$

- Rationale: reduce variance by constraining the model.
- Some possible forms for the penalty term:
 - L_2 arising from Gaussian $p(\mathbf{w})$: $\sum_j w_j^2$.
 - L_1 arising from Laplacian $p(\mathbf{w})$: $\sum_j |w_j|$.
 - Can define many other types of penalty terms...
- The regularization parameter λ determines the strength of the penalty contribution to the objective.

Plan for today

- Regularization in regression.
- A brief survey of where we are and what we have learned.
- Large margin classifiers.

Shrinkage / Ridge regression

- We can impose penalty on \mathbf{w} in a way similar to LR.
- First, let's assume Gaussian noise model, and L_2 regularization. The penalized log-likelihood is:

$$-\sum_{i=1}^N (y_i - \mathbf{w}^T \mathbf{x}_i)^2 - \lambda \sum_{j=1}^d w_j^2$$

- This is known in statistics as *ridge regression*, or *parameter shrinkage*.
- The solution (done in PS3):

$$\hat{\mathbf{w}}_{ridge} = (\lambda \mathbf{I} + \mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

- I.e., still a unique maximum obtained in closed-form!

Lasso regression

- The L_1 -penalized log-likelihood under Gaussian noise model:

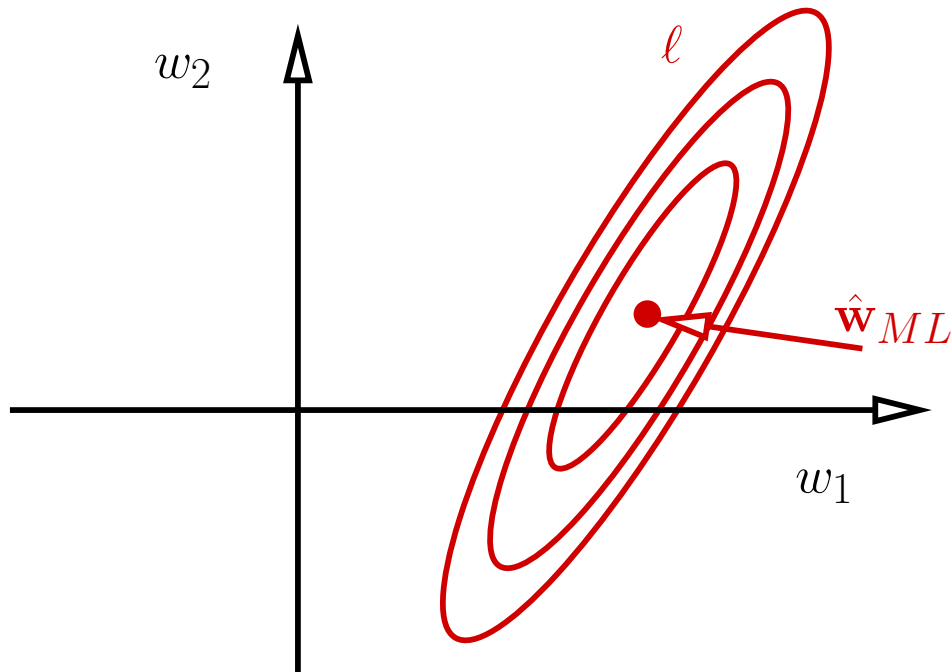
$$-\sum_{i=1}^N (y_i - \mathbf{w}^T \mathbf{x}_i)^2 - \lambda \sum_{j=1}^d |w_j|$$

- This is still concave (i.e. unique maximum), but unfortunately neither closed-form solution nor gradient descent will do the trick.
 - the objective is not “smooth”.
- Why is it called “lasso”?

Lasso vs. ridge: geometry of error surfaces

- An equivalent formulation for L_p regularization: constrained maximization

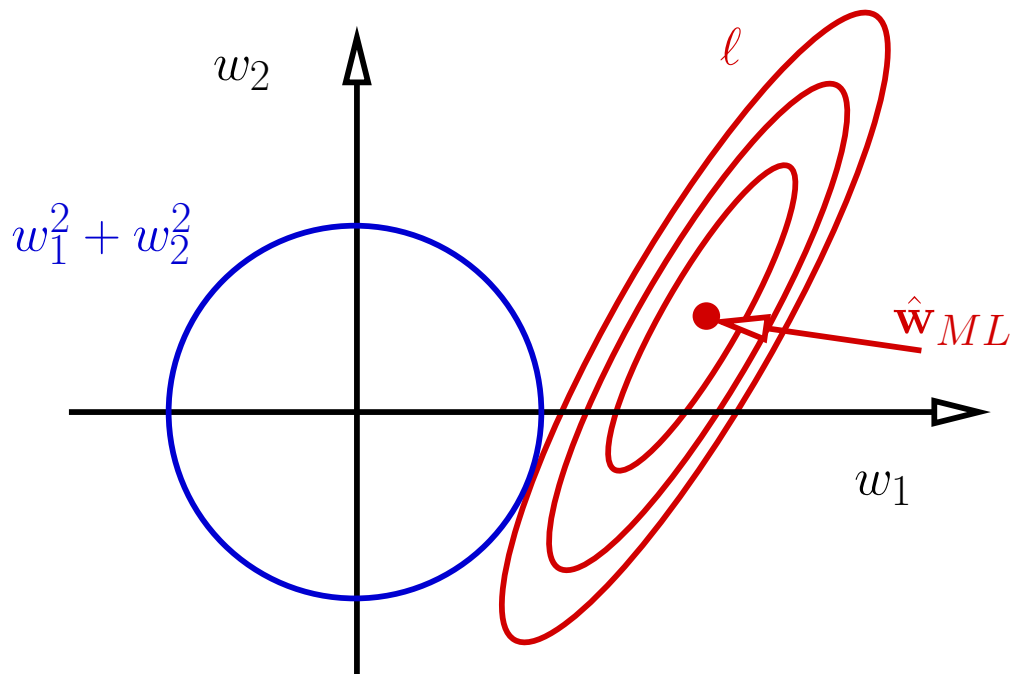
$$\hat{\mathbf{w}} = \underset{\mathbf{w}: \sum_{j=1}^d |w_j|^p \leq \beta}{\operatorname{argmax}} - \sum_{i=1}^N (y_i - \mathbf{w}^T \mathbf{x}_i)^2.$$



Lasso vs. ridge: geometry of error surfaces

- An equivalent formulation for L_p regularization: constrained maximization

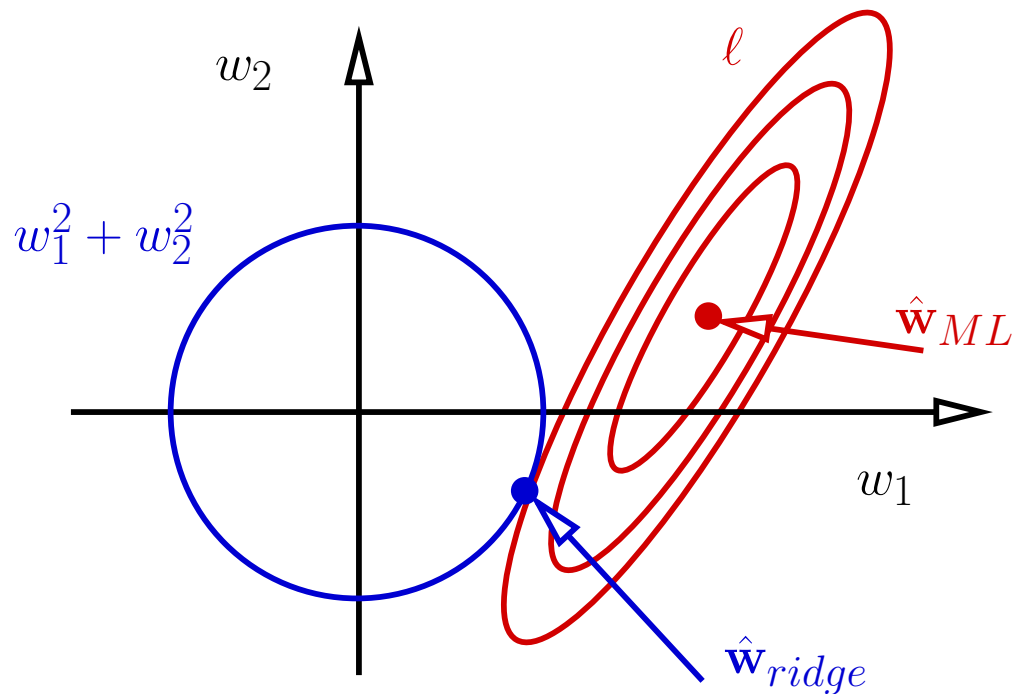
$$\hat{\mathbf{w}} = \underset{\mathbf{w}: \sum_{j=1}^d |w_j|^p \leq \beta}{\operatorname{argmax}} - \sum_{i=1}^N (y_i - \mathbf{w}^T \mathbf{x}_i)^2.$$



Lasso vs. ridge: geometry of error surfaces

- An equivalent formulation for L_p regularization: constrained maximization

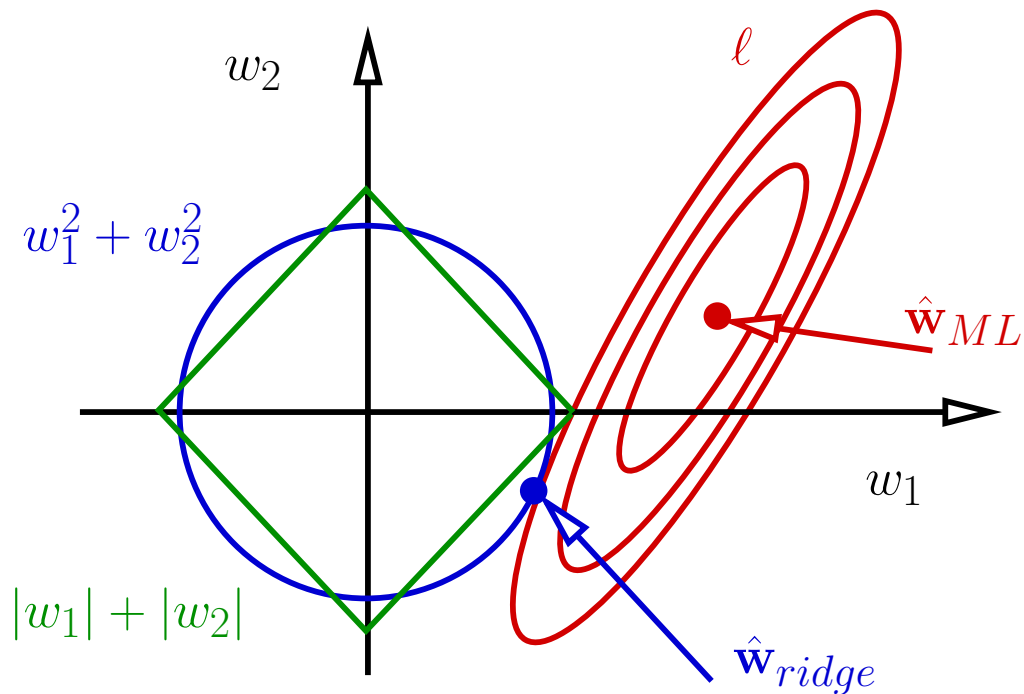
$$\hat{\mathbf{w}} = \underset{\mathbf{w}: \sum_{j=1}^d |w_j|^p \leq \beta}{\operatorname{argmax}} - \sum_{i=1}^N (y_i - \mathbf{w}^T \mathbf{x}_i)^2.$$



Lasso vs. ridge: geometry of error surfaces

- An equivalent formulation for L_p regularization: constrained maximization

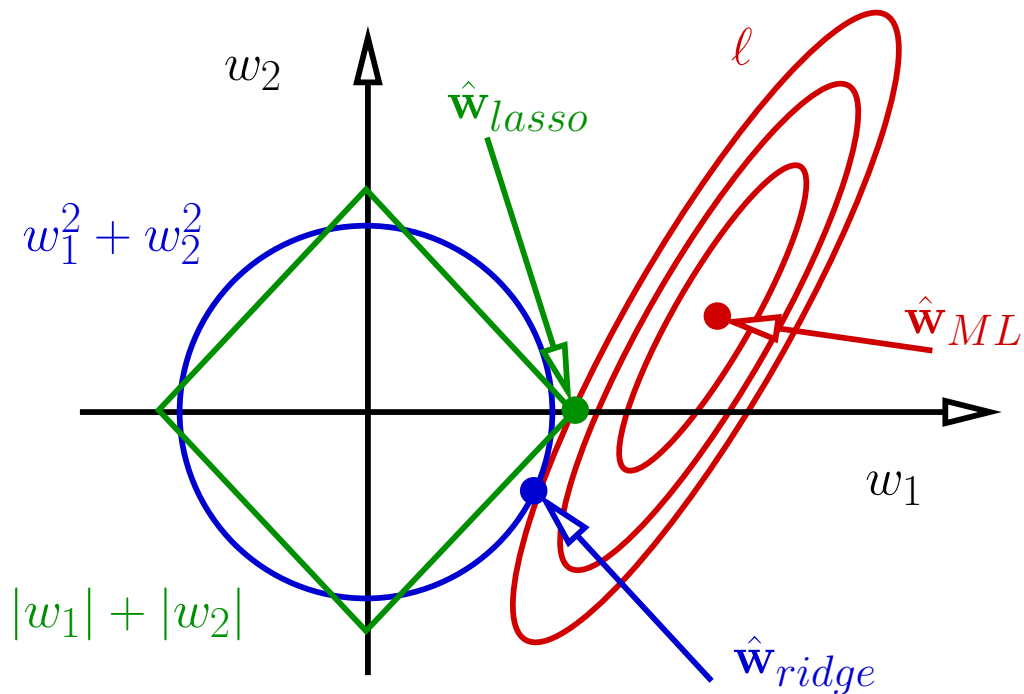
$$\hat{\mathbf{w}} = \underset{\mathbf{w}: \sum_{j=1}^d |w_j|^p \leq \beta}{\operatorname{argmax}} - \sum_{i=1}^N (y_i - \mathbf{w}^T \mathbf{x}_i)^2.$$



Lasso vs. ridge: geometry of error surfaces

- An equivalent formulation for L_p regularization: constrained maximization

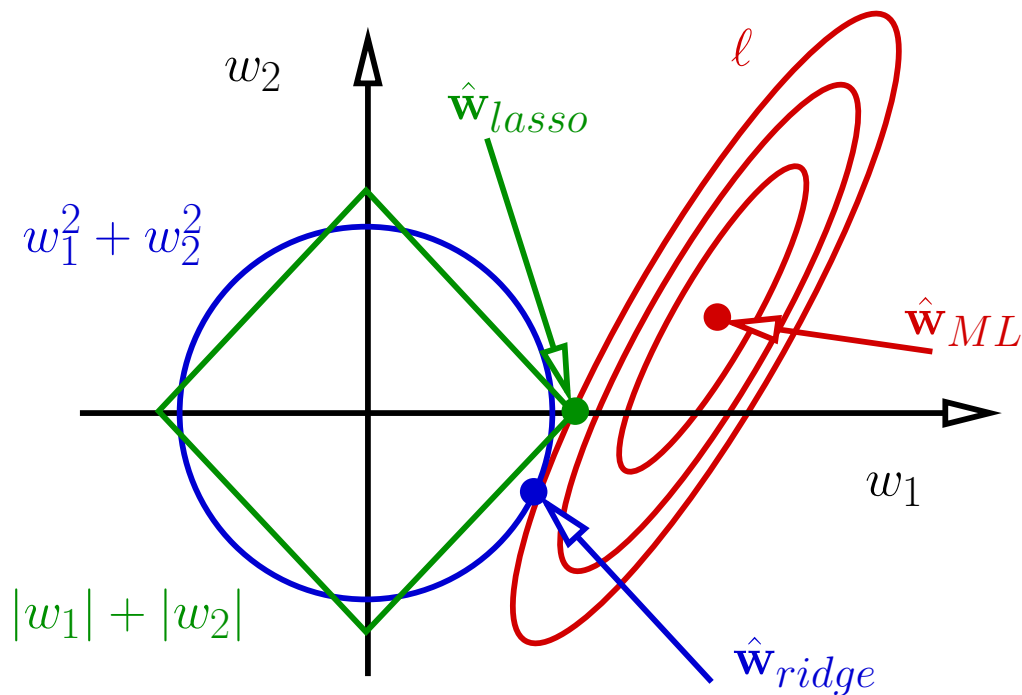
$$\hat{\mathbf{w}} = \underset{\mathbf{w}: \sum_{j=1}^d |w_j|^p \leq \beta}{\operatorname{argmax}} - \sum_{i=1}^N (y_i - \mathbf{w}^T \mathbf{x}_i)^2.$$



Lasso vs. ridge: geometry of error surfaces

- An equivalent formulation for L_p regularization: constrained maximization

$$\hat{\mathbf{w}} = \underset{\mathbf{w}: \sum_{j=1}^d |w_j|^p \leq \beta}{\operatorname{argmax}} - \sum_{i=1}^N (y_i - \mathbf{w}^T \mathbf{x}_i)^2.$$

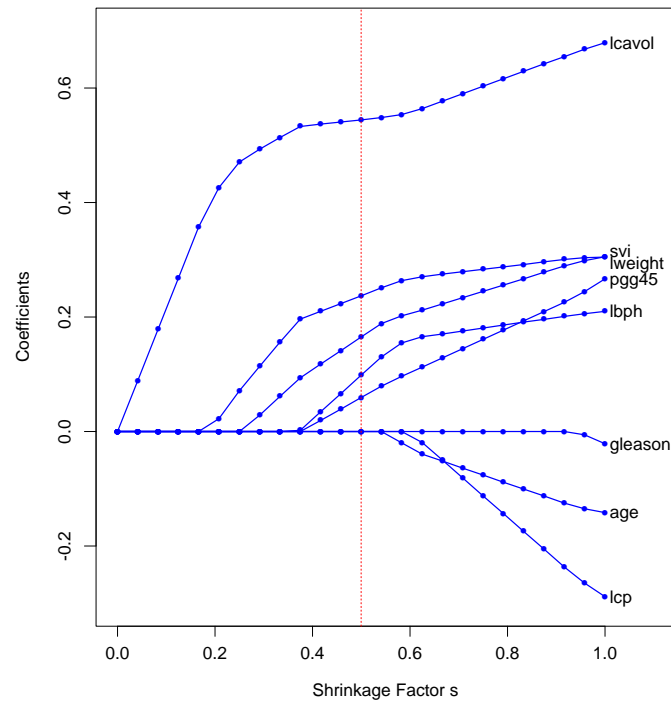
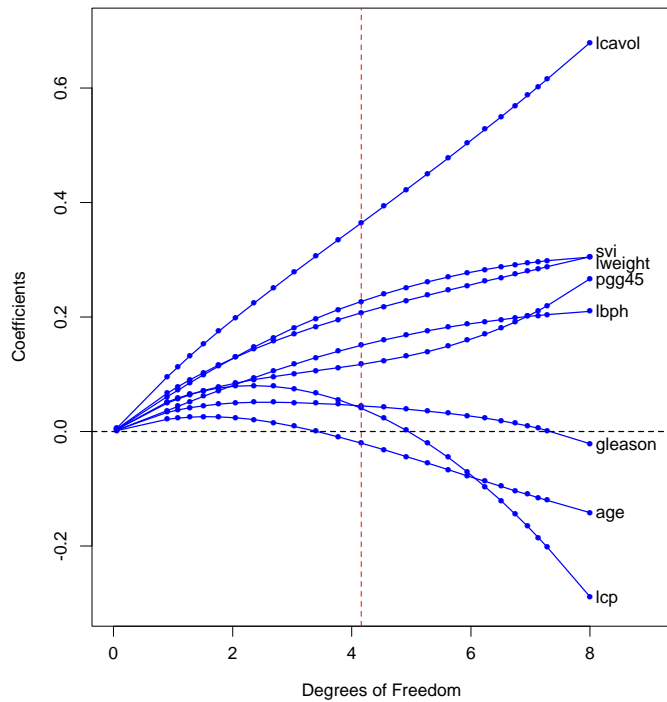


- With sufficiently large λ , lasso leads to *sparsity*.
- Must explicitly solve the above optimization problem – e.g., using Lagrange multipliers.

Example: lasso vs. ridge

From HTF: prostate data

Red lines: choice of λ by 10-fold CV.



What have we seen so far

- Fundamental concepts:
 - Learning via empirical loss minimization
 - Bias-variance tradeoff
 - Overfitting and generalization
 - Model selection: cross-validation.
 - Estimation: “frequentist” (ML) and “Bayesian” (MAP).
- A number of models and learning algorithms

Algorithms for supervised learning

Regression

- Generalized linear regression models.

Classification

- Generative models:
 - Gaussian class-conditionals \Rightarrow linear or quadratic discriminant analysis
 - Naïve Bayes classifiers, with Bernoulli marginal class-conditionals.
- Discriminative models
 - Logistic regression and softmax.
 - Fisher's LDA

Some rules of thumb

- Smaller data sets \Rightarrow need to worry more about variance and overfitting.
- Simpler models \Rightarrow may suffer from bias (but less likely to overfit).
 - Simpler = more restricted: fewer parameters or constraints on parameters (penalty, parameters “tied up” etc.)
- In many cases a model/algorithm which is optimal under some assumptions that are clearly violated in the data may still work very well:
 - Fisher’s LDA, Naïve Bayes, Gaussian class model (LDA/QDA),...

Discriminative versus generative models

Main distinction:

- Generative: model prior $p(y)$ and class-conditional $p(\mathbf{x} | y)$.
- Discriminative: model posterior $p(y | \mathbf{x})$ directly.
- The ultimate criterion: choose one that works better on test set / CV.
 - If you have a good reason to believe the generative model, go for it (but beware insufficient data!)
 - Anecdote: if the classes *are* Gaussian, but you ignore that and use linear logistic regression, you are 30% less efficient.
- Often discriminative models happen to have fewer parameters – an advantage on small data sets.

Discriminative classification

- We are still in the realm of linear classification

$$\hat{y}(\mathbf{x}) = \text{sign}(w_0 + \mathbf{w}^T \mathbf{x}).$$

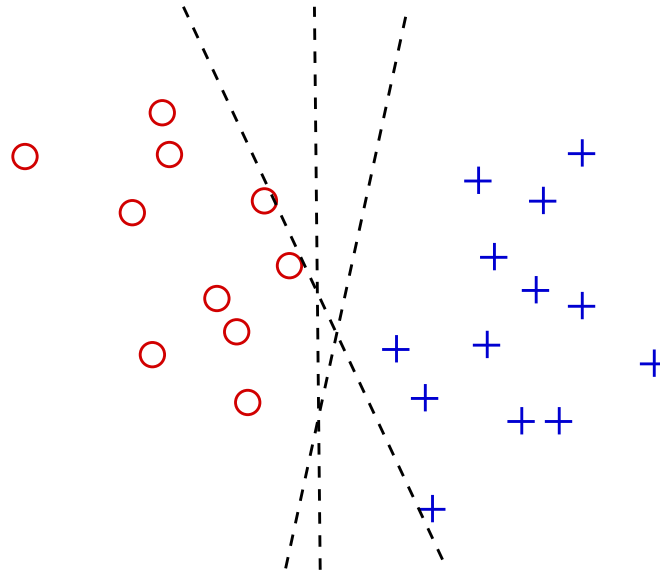
- Our eventual objective is to minimize expected 0/1 risk:

$$E_{y,\mathbf{x}} [L(\hat{y}(\mathbf{x}), y)].$$

- No probabilities are associated with the predictions \hat{y} in this formulation; we need to produce a “hard” class assignment for the test \mathbf{x} .

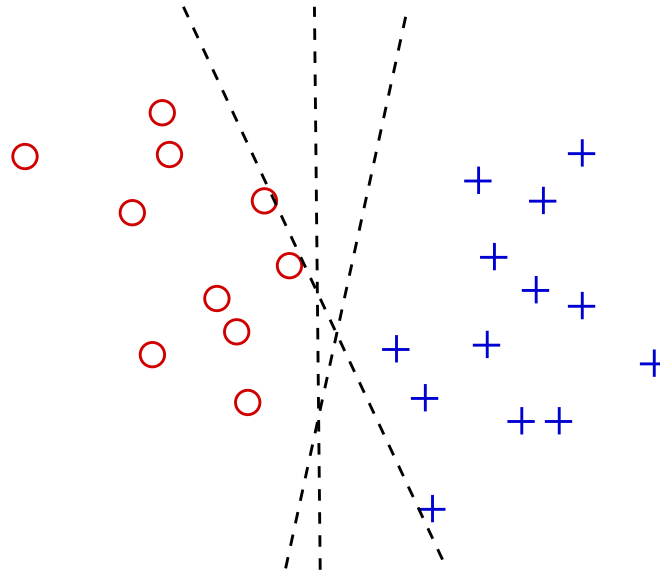
Two-class, linearly separable data

- Which linear decision boundary is better?



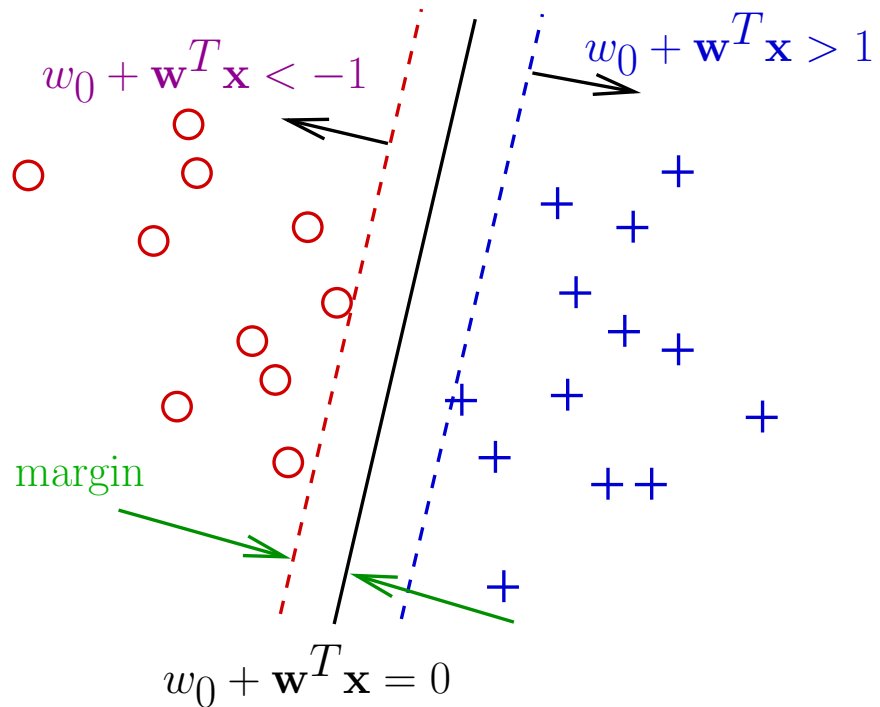
Two-class, linearly separable data

- Which linear decision boundary is better?



- A possible criterion: the boundary that maximizes the separation between classes.

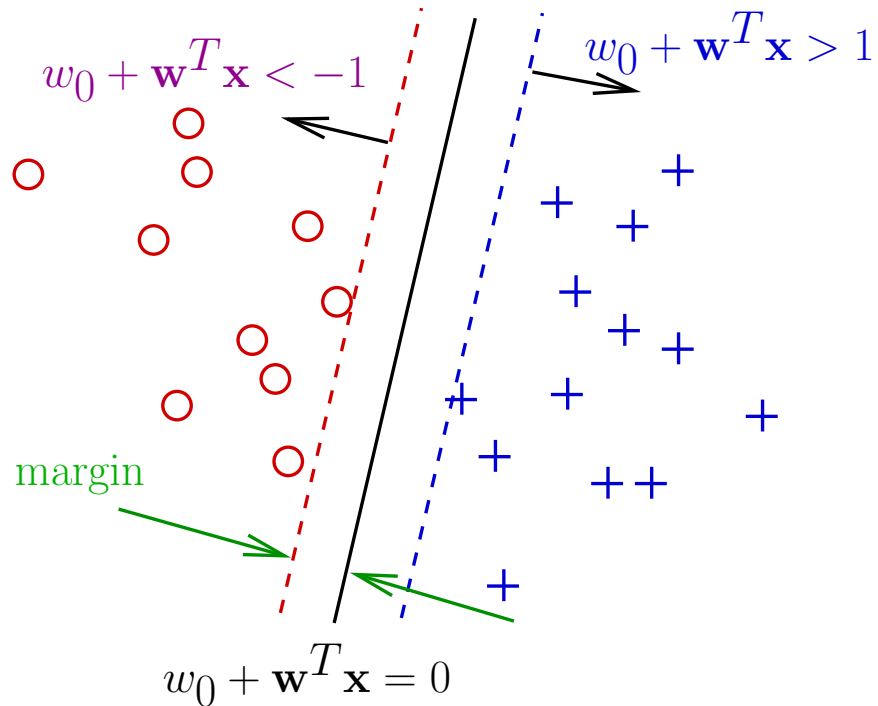
The classification margin



- Since the data are separable, we can find \mathbf{w} such that

$$\forall i = 1, \dots, N \quad y_i(w_0 + \mathbf{w}^T \mathbf{x}_i) > 0.$$

The classification margin



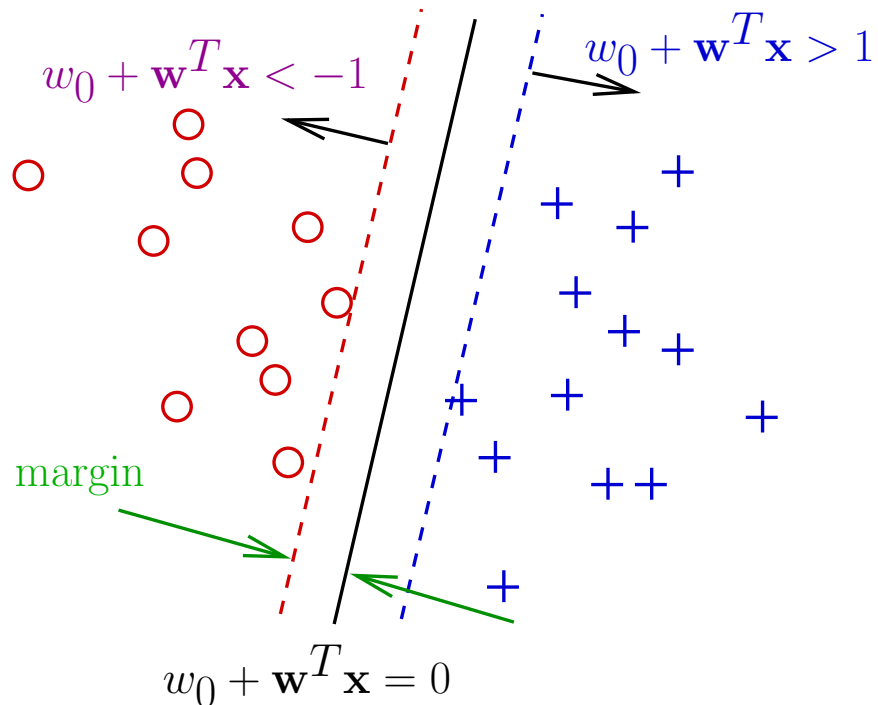
- Since the data are separable, we can find \mathbf{w} such that

$$\forall i = 1, \dots, N \quad y_i(w_0 + \mathbf{w}^T \mathbf{x}_i) > 0.$$

- We can even guarantee (by increasing $\|\mathbf{w}\|$ if necessary)

$$\forall i = 1, \dots, N \quad y_i(w_0 + \mathbf{w}^T \mathbf{x}_i) \geq 1.$$

The classification margin



- Since the data are separable, we can find \mathbf{w} such that

$$\forall i = 1, \dots, N \quad y_i(w_0 + \mathbf{w}^T \mathbf{x}_i) > 0.$$

- We can even guarantee (by increasing $\|\mathbf{w}\|$ if necessary)

$$\forall i = 1, \dots, N \quad y_i(w_0 + \mathbf{w}^T \mathbf{x}_i) \geq 1.$$

- $\min_i y_i(w_0 + \mathbf{w}^T \mathbf{x}_i)$ is the smallest distance from \mathbf{x}_i to the boundary (half the separation between classes).
- We will refer to it as the *margin*.

Max-margin boundary

- Can we just state that we want

$$\hat{\mathbf{w}} = \operatorname{argmax}_{\mathbf{w}} \min_i y_i (w_0 + \mathbf{w}^T \mathbf{x}_i)?$$

Max-margin boundary

- Can we just state that we want

$$\hat{\mathbf{w}} = \operatorname{argmax}_{\mathbf{w}} \min_i y_i (w_0 + \mathbf{w}^T \mathbf{x}_i)?$$

- Same kind of problem we have seen with LR: when data are separable the margin is unbounded as $\|\mathbf{w}\| \rightarrow \infty$.
- Suppose $y = 1$, and $\|\mathbf{w}\| = 1$. Let $w_0 + \mathbf{w}^T \mathbf{x} = c$. Then,

$$\alpha w_0 + (\alpha \cdot \mathbf{w})^T \mathbf{w} = \alpha (w_0 + \mathbf{w}^T \mathbf{x}) = \alpha c,$$

Max-margin boundary

- Can we just state that we want

$$\hat{\mathbf{w}} = \operatorname{argmax}_{\mathbf{w}} \min_i y_i (w_0 + \mathbf{w}^T \mathbf{x}_i)?$$

- Same kind of problem we have seen with LR: when data are separable the margin is unbounded as $\|\mathbf{w}\| \rightarrow \infty$.
- Suppose $y = 1$, and $\|\mathbf{w}\| = 1$. Let $w_0 + \mathbf{w}^T \mathbf{x} = c$. Then,

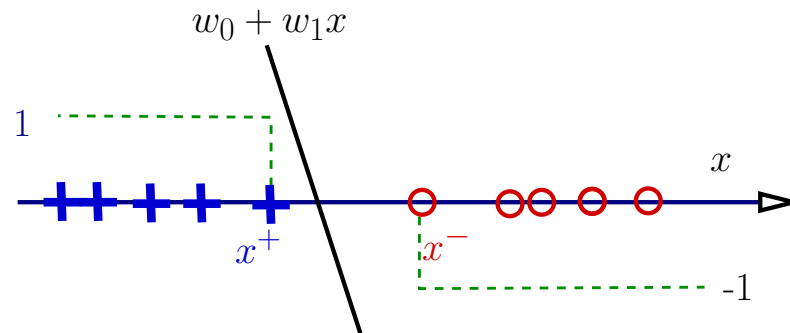
$$\alpha w_0 + (\alpha \cdot \mathbf{w})^T \mathbf{w} = \alpha (w_0 + \mathbf{w}^T \mathbf{x}) = \alpha c,$$

i.e. we can achieve arbitrarily wide margin with the same classification boundary.

- We could require $\|\mathbf{w}\| = 1$.

Fixed margin solution

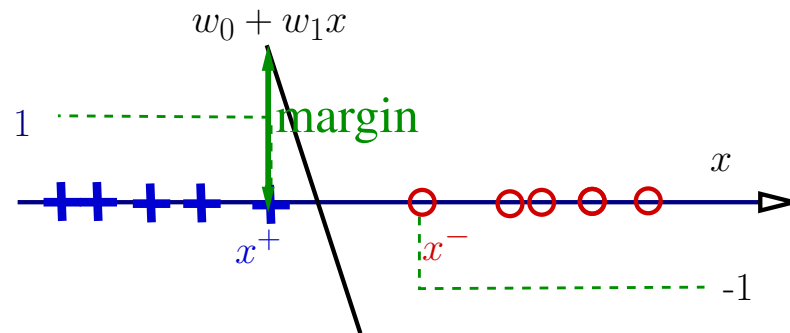
- A more convenient solution: require *fixed* margin of, say, 1.
- Of all \mathbf{w} that achieve such margin, choose the smallest one.
 - This imposes a unique (equivalent) solution!
- The margin constraints, graphically:



$$\begin{aligned} 1 \cdot (w_0 + w_1 x_i) - 1 &\geq 0, & y_i &= 1 \\ -1 \cdot (w_0 + w_1 x_i) - 1 &\geq 0, & y_i &= -1. \end{aligned}$$

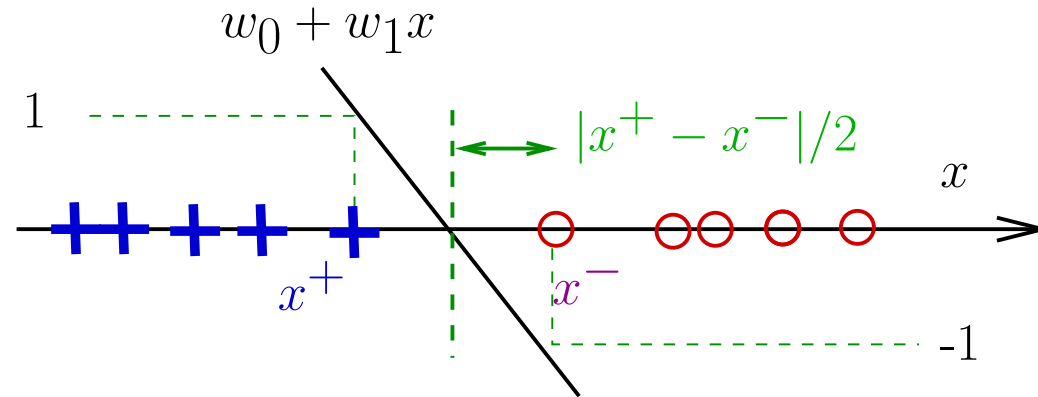
Fixed margin solution

- A more convenient solution: require *fixed* margin of, say, 1.
- Of all \mathbf{w} that achieve such margin, choose the smallest one.
 - This imposes a unique (equivalent) solution!
- The margin constraints, graphically:



$$\begin{aligned} 1 \cdot (w_0 + w_1 x_i) - 1 &\geq 0, & y_i &= 1 \\ -1 \cdot (w_0 + w_1 x_i) - 1 &\geq 0, & y_i &= -1. \end{aligned}$$

Margin vs. slope



- Separation is maximal when the line passes through $(x^+ + x^-)/2$.
 - The maximum margin is 1;
- the margin is *inversely proportional* to the slope $|w_1|$;
- The optimal boundary is achieved with

$$|w_1| = 2/|x^+ - x^-|.$$

Margin and regularization

- In general d -dimensional case, we solve the regularization problem:

$$\text{minimize} \quad \frac{1}{2} \|\mathbf{w}\|^2 = \frac{1}{2} \sum_{j=1}^d w_j^2,$$

subject to the margin constraint

$$\forall i, \quad y_i(w_0 + \mathbf{w}^T \mathbf{x}_i) - 1 \geq 0.$$

- This produces margin of exactly 1 (why?)
- Again, the solution is expressed in terms of only a subset of examples.
 - These are the *support vectors*.

Next time

Support Vector Machines.