In Memoriam: Marvin Minsky

Will robots inherit the earth? Yes, but they will be our children

Lecture 5: More on Kernels. SVM regression and classification

- How to tell if a function is a kernel
- SVM regression
- SVM classification

The "kernel trick"

- Recall: kernel functions are ways of expressing dot-products in some feature space:

$$K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{z})$$

- In many cases, K can be computed in time that depends on the size of the inputs $\mathbf{x}$ not the size of the feature space $\phi\mathbf{x}$
- If we work with a "dual" representation of the learning algorithm, we do not actually have to compute the feature mapping ϕ. We just have to compute the similarity K.

Polynomial kernels

- More generally, $K(\mathbf{x}, \mathbf{z}) = (\mathbf{x} \cdot \mathbf{z})^d$ is a kernel, for any positive integer d:

$$K(\mathbf{x}, \mathbf{z}) = \left(\sum_{i=1}^{n} x_i z_i \right)^d$$

- If we expanded the sum above in the obvious way, we get n^d terms (i.e. feature expansion)
- Terms are monomials (products of x_i) with total power equal to d.
- *Curse of dimensionality:* it is very expensive both to optimize and to predict in primal form
- However, *evaluating the dot-product of any two feature vectors can be done using K in $O(n)$*!

Some other (fairly generic) kernel functions

- $K(\mathbf{x}, \mathbf{z}) = (1 + \mathbf{x} \cdot \mathbf{z})^d$ – feature expansion has all monomial terms of $\leq d$ total power.
- Radial basis/Gaussian kernel:

$$K(\mathbf{x}, \mathbf{z}) = \exp(-\|\mathbf{x} - \mathbf{z}\|^2 / 2\sigma^2)$$

  The kernel has an infinite-dimensional feature expansion, but dot-products can still be computed in $O(n)$!
- Sigmoidal kernel:

$$K(\mathbf{x}, \mathbf{z}) = \tanh(c_1 \mathbf{x} \cdot \mathbf{z} + c_2)$$

Recall: Dual-view regression

- By re-writing the parameter vector as a linear combination of instances and solving, we get:

$$\mathbf{a} = (\mathbf{K} + \lambda \mathbf{I}_m)^{-1}\mathbf{y}$$

- The *feature mapping is not needed* either to learn or to make predictions!
- This approach is useful if the feature space is very large

Making predictions in the dual view

- For a new input $\mathbf{x}$, the prediction is:

$$h(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) = \mathbf{a}^T \mathbf{\Phi} \phi(\mathbf{x}) = \mathbf{k}(\mathbf{x})^T (\mathbf{K} + \lambda \mathbf{I}_m)^{-1} \mathbf{y}$$

  where $\mathbf{k}(\mathbf{x})$ is an m-dimensional vector, with the ith element equal to $K(\mathbf{x}, \mathbf{x}_i)$

- That is, the ith element has the similarity of the input to the ith instance
- The features are not needed for this step either!
- This is a *non-parametric* representation - its size scales with the number of instances.

Regularization in the dual view

- We want to penalize the function we are trying to estimate (to keep it simple)
- Assume this is part of a reproducing kernel Hilbert space (Doina will post extra notes for those interested in this)
- We want to minimize:

$$J(h) = \frac{1}{2}\sum_{i=1}^{n}(y_i - h(\mathbf{x}_i))^2 + \frac{\lambda}{2}||h||^2_{\mathcal{H}}$$

- If we put a Gaussian prior on h, and solve, we obtain *Gaussian process regression*

Logistic regression

- The output of a logistic regression predictor is:

$$h_{\mathbf{w}}(\mathbf{x}) = \frac{1}{1 + e^{\mathbf{w}^T \phi(\mathbf{x}) + w_0}}$$

- Again, we can define the weights in terms of support vectors: $\mathbf{w} = \sum_{i=1}^{m} \alpha_i \phi(\mathbf{x}_i)$
- The prediction can now be computed as:

$$h(\mathbf{x}) = \frac{1}{1 + e^{\sum_{i=1}^{m} \alpha_i K(\mathbf{x}_i, \mathbf{x}) + w_0}}$$

- α_i are the new parameters (one per instance) and can be derived using gradient descent

Kernels

- A lot of current research has to do with defining new kernels functions, suitable to particular tasks / kinds of input objects
- Many kernels are available:
  - Information diffusion kernels (Lafferty and Lebanon, 2002)
  - Diffusion kernels on graphs (Kondor and Jebara 2003)
  - String kernels for text classification (Lodhi et al, 2002)
  - String kernels for protein classification (e.g., Leslie et al, 2002)

  ... and others!

Example: String kernels

- Very important for DNA matching, text classification, ...
- Example: in DNA matching, we use a sliding window of length k over the two strings that we want to compare
- The window is of a given size, and inside we can do various things:
  - Count exact matches
  - Weigh mismatches based on how bad they are
  - Count certain markers, e.g. AGT
- The kernel is the sum of these similarities over the two sequences
- How do we prove this is a kernel?

Establishing "kernelhood"

- Suppose someone hands you a function K. How do you know that it is a kernel?
- More precisely, given a function $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, under what conditions can $K(\mathbf{x}, \mathbf{z})$ be written as a dot product $\phi(\mathbf{x}) \cdot \phi(\mathbf{z})$ for some feature mapping ϕ?
- We want a general recipe, which does not require explicitly defining ϕ every time

Kernel matrix

- Suppose we have an arbitrary set of input vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_m$
- The *kernel matrix (or Gram matrix)* K corresponding to kernel function K is an $m \times m$ matrix such that $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ (notation is overloaded on purpose).
- What properties does the kernel matrix K have?
- Claims:
  1. K is symmetric
  2. K is positive semidefinite
- Note that these claims are consistent with the intuition that K is a "similarity" measure (and will be true regardless of the data)

Proving the first claim

If K is a valid kernel, then the kernel matrix is symmetric

$$K_{ij} = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) = \phi(\mathbf{x}_j) \cdot \phi(\mathbf{x}_i) = K_{ji}$$

Proving the second claim

If K is a valid kernel, then the kernel matrix is positive semidefinite

Proof: Consider an arbitrary vector $\mathbf{z}$

$$\begin{aligned}
\mathbf{z}^T K \mathbf{z} &= \sum_i \sum_j z_i K_{ij} z_j = \sum_i \sum_j z_i \left(\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)\right) z_j \\
&= \sum_i \sum_j z_i \left(\sum_k \phi_k(\mathbf{x}_i) \phi_k(\mathbf{x}_j) \right) z_j \\
&= \sum_k \sum_i \sum_j z_i \phi_k(\mathbf{x}_i) \phi_k(\mathbf{x}_j) z_j \\
&= \sum_k \left(\sum_i z_i \phi_k(\mathbf{x}_i) \right)^2 \geq 0
\end{aligned}$$

Mercer's theorem

- We have shown that if K is a kernel function, then for any data set, the corresponding kernel matrix K defined such that $K_{ij} = K(\mathbf{x_i}, \mathbf{x_j})$ is symmetric and positive semidefinite
- Mercer's theorem states that the reverse is also true:
  Given a function $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, *$K$ is a kernel if and only if, for any data set, the corresponding kernel matrix is symmetric and positive semidefinite*
- The reverse direction of the proof is much harder (see e.g. Vapnik's book for details)
- This result gives us a way to check if a given function is a kernel, by checking these two properties of its kernel matrix.
- Kernels can also be obtained by combining other kernels, or by learning from data
- Kernel learning may suffer from overfitting (kernel matrix close to diagonal)

Support Vector Regression

- In regression problems, so far we have been trying to minimize mean-squared error:

$$\sum_i (y_i - (\mathbf{w} \cdot \mathbf{x_i} + w_0))^2$$

- In SVM regression, we will be interested instead in minimizing absolute error:

$$\sum_i |y_i - (\mathbf{w} \cdot \mathbf{x_i} + w_0)|$$

- This is more *robust to outliers* than the squared loss
- But we cannot require that all points be approximated correctly (overfitting!)

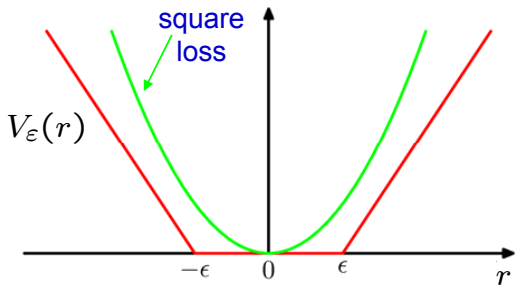
Loss function for support vector regression

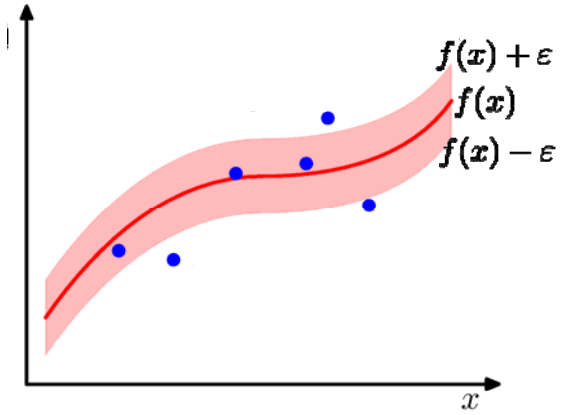
In order to allow for misclassifications in SVM regression (and have robustness to noise), we use the *ϵ-insensitive loss*:

$$J_\epsilon = \sum_{i=1}^{m} J_\epsilon(\mathbf{x_i}), \text{ where}$$

$$J_\epsilon(\mathbf{x_i}) = \begin{cases} 0 & \text{if } |y_i - (\mathbf{w} \cdot \mathbf{x_i} + w_0)| \leq \epsilon \\ |y_i - (\mathbf{w} \cdot \mathbf{x_i} + w_0)| - \epsilon & \text{otherwise} \end{cases}$$

cost is zero inside epsilon "tube"





1

Solving SVM regression

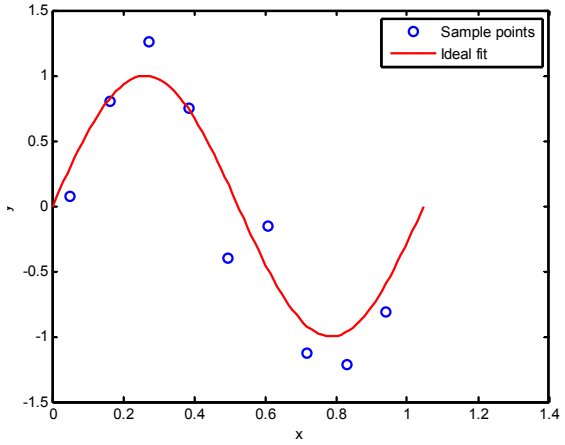
- We introduce slack variables, ξ_i^+, ξ_i^- to account for errors outside the tolerance area
- We need two kinds of variables to account for both positive and negative errors

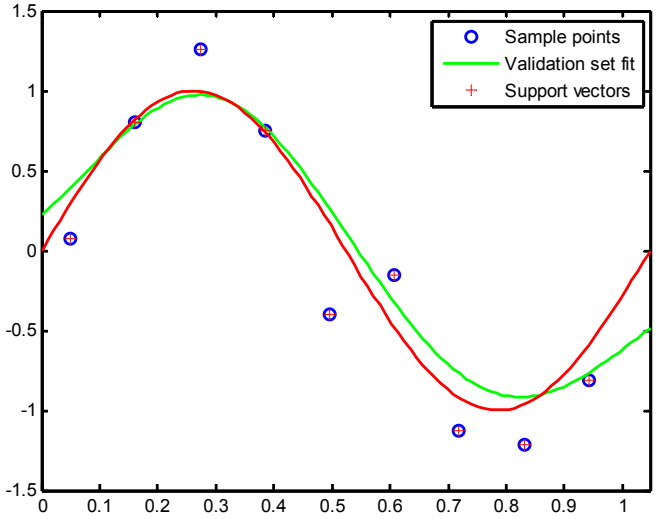
The optimization problem

$$
\begin{array}{rl}
\text{min} & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i^+ + \xi_i^-) \\
\text{w.r.t.} & \mathbf{w}, w_0, \xi_i^+, \xi_i^- \\
\text{s.t.} & y_i - (\mathbf{w}\cdot\mathbf{x}_i + w_0) \leq \epsilon + \xi_i^+ \\
& y_i - (\mathbf{w}\cdot\mathbf{x}_i + w_0) \geq -\epsilon - \xi_i^- \\
& \xi_i^+, \xi_i^- \geq 0
\end{array}
$$

- Like before, we can write the Lagrangian and solve the dual form of the problem
- Kernels can be used as before to get non-linear functions

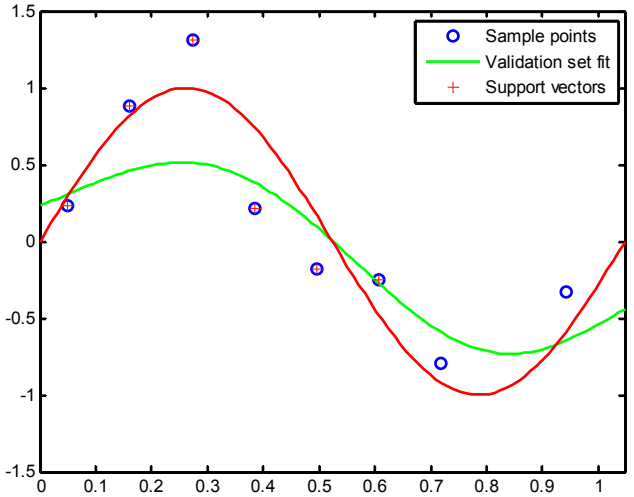
Effect of ϵ





epsilon = 0.01

epsilon = 0.8



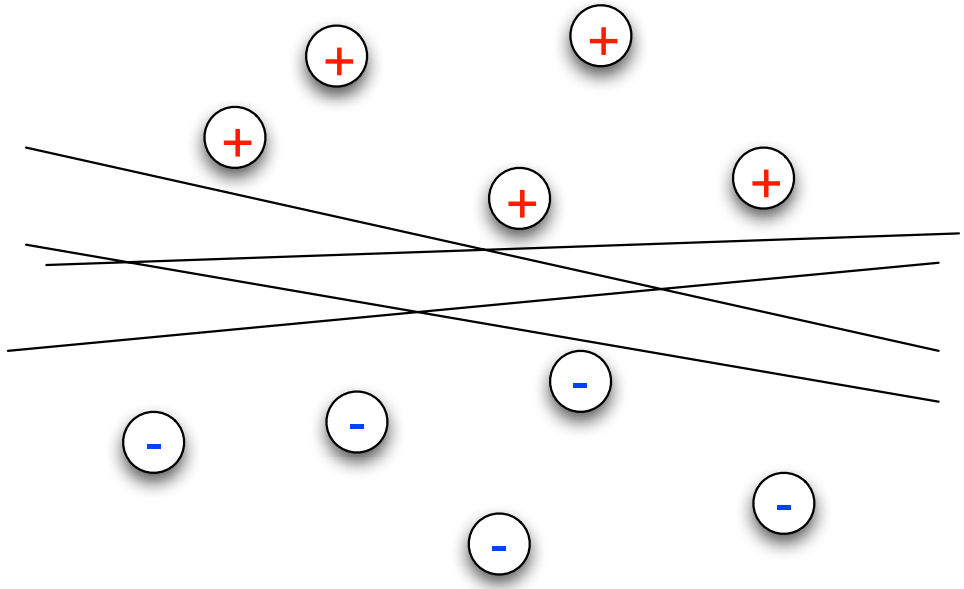
[2]

- As ϵ increases, the function is allowed to move away from the data points, the number of support vectors decreases and the fit gets worse

[2] Zisserman course notes

Binary classification revisited

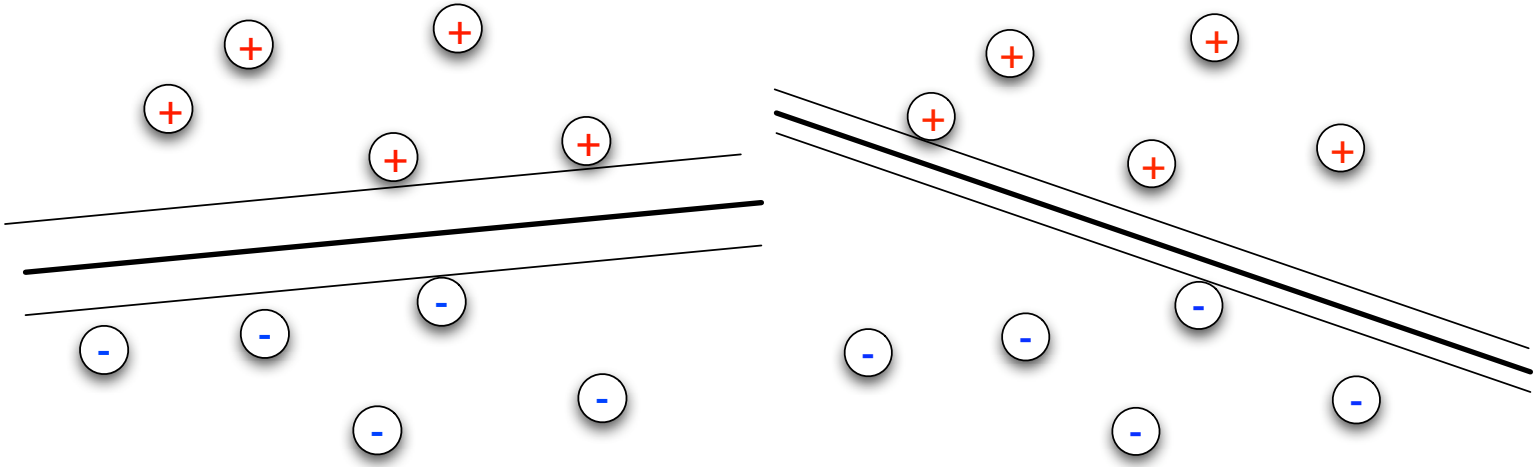
- Consider a linearly separable binary classification data set $\{\mathbf{x}_i, y_i\}_{i=1}^m$.
- There is an infinite number of hyperplanes that separate the classes:



- Which plane is best?
- Relatedly, for a given plane, for which points should we be most confident in the classification?

The margin, and linear SVMs

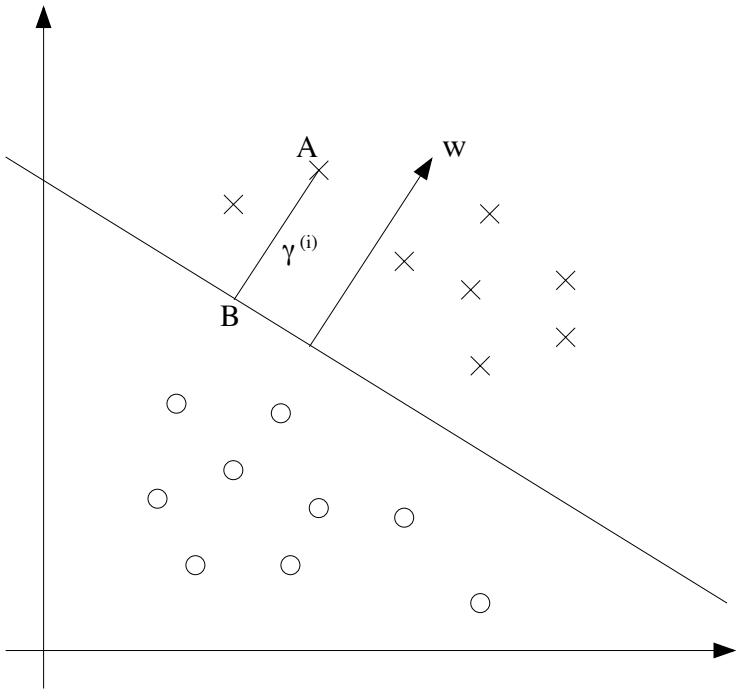
- For a given separating hyperplane, the *margin* is two times the (Euclidean) distance from the hyperplane to the nearest training example.



- It is the width of the "strip" around the decision boundary containing no training examples.
- A linear SVM is a perceptron for which we choose $\mathbf{w}, w_0$ so that margin is maximized

Distance to the decision boundary

- Suppose we have a decision boundary that separates the data.



- Let γ_i be the distance from instance $\mathbf{x}_i$ to the decision boundary.
- How can we write γ_i in term of $\mathbf{x}_i, y_i, \mathbf{w}, w_0$?

Distance to the decision boundary (II)

- The vector $\mathbf{w}$ is normal to the decision boundary. Thus, $\frac{\mathbf{w}}{||\mathbf{w}||}$ is the unit normal.
- The vector from the B to A is $\gamma_i \frac{\mathbf{w}}{||\mathbf{w}||}$.
- B, the point on the decision boundary nearest $\mathbf{x}_i$, is $\mathbf{x}_i - \gamma_i \frac{\mathbf{w}}{||\mathbf{w}||}$.
- As B is on the decision boundary,

$$\mathbf{w} \cdot \left(\mathbf{x}_i - \gamma_i \frac{\mathbf{w}}{||\mathbf{w}||} \right) + w_0 = 0$$

- Solving for γ_i yields, for a positive example:

$$\gamma_i = \frac{\mathbf{w}}{||\mathbf{w}||} \cdot \mathbf{x}_i + \frac{w_0}{||\mathbf{w}||}$$

The margin

- The *margin of the hyperplane* is $2M$, where $M = \min_i \gamma_i$
- The most direct statement of the problem of finding a maximum margin separating hyperplane is thus

$$\begin{aligned} & \max_{\mathbf{w}, w_0} \min_i \gamma_i \\ \equiv \quad & \max_{\mathbf{w}, w_0} \min_i y_i \left(\frac{\mathbf{w}}{||\mathbf{w}||} \cdot \mathbf{x}_i + \frac{w_0}{||\mathbf{w}||} \right) \end{aligned}$$

- This turns out to be inconvenient for optimization, however. . .

Treating the γ_i as constraints

- From the definition of margin, we have:

$$M \leq \gamma_i = y_i \left(\frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot \mathbf{x_i} + \frac{w_0}{\|\mathbf{w}\|} \right) \qquad \forall i$$

- This suggests:

$$\begin{array}{rl} \text{maximize} & M \\ \text{with respect to} & \mathbf{w}, w_0 \\ \text{subject to} & \mathbf{y}_i \left(\frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot \mathbf{x}_i + \frac{w_0}{\|\mathbf{w}\|} \right) \geq M \text{ for all } i \end{array}$$

- Problems:
  - $\mathbf{w}$ appears nonlinearly in the constraints.
  - This problem is underconstrained. If $(\mathbf{w}, w_0, M)$ is an optimal solution, then so is $(\beta\mathbf{w}, \beta w_0, M)$ for any $\beta > 0$.

Adding a constraint

- Let's try adding the constraint that $\|\mathbf{w}\|M = 1$.
- This allows us to rewrite the objective function and constraints as:

$$\begin{aligned} \text{min} \quad & \|\mathbf{w}\| \\ \text{w.r.t.} \quad & \mathbf{w}, w_0 \\ \text{s.t.} \quad & y_i(\mathbf{w} \cdot \mathbf{x}_i + w_0) \geq 1 \end{aligned}$$

- This is really nice because the constraints are linear.
- The objective function $\|\mathbf{w}\|$ is still a bit awkward.

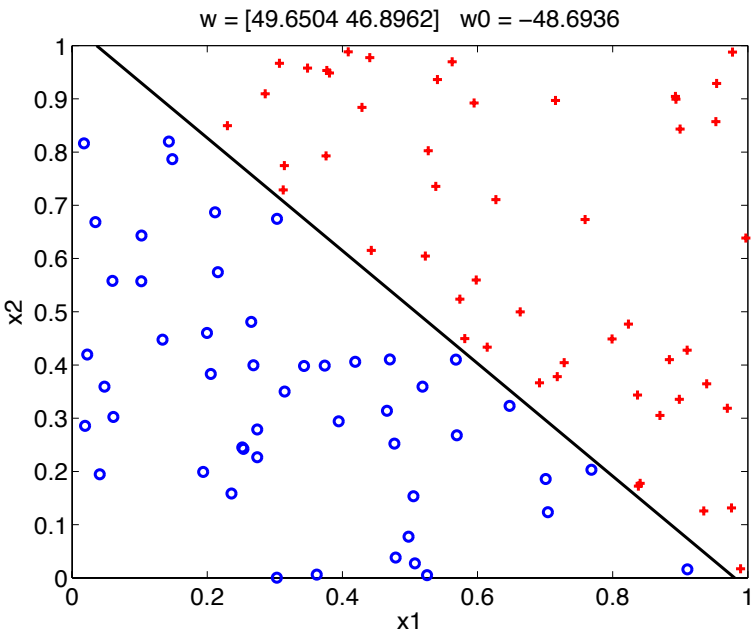
Final formulation

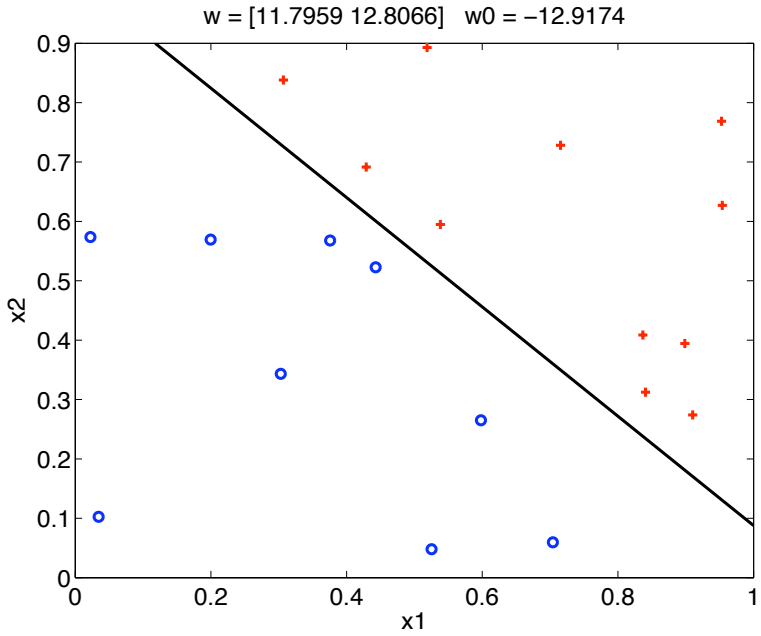
- Let's maximize $\|\mathbf{w}\|^2$ instead of $\|\mathbf{w}\|$.
  (Taking the square is a monotone transformation, as $\|\mathbf{w}\|$ is postive, so this doesn't change the optimal solution.)
- This gets us to:

$$\begin{aligned} \text{min} \quad & \|\mathbf{w}\|^2 \\ \text{w.r.t.} \quad & \mathbf{w}, w_0 \\ \text{s.t.} \quad & y_i(\mathbf{w} \cdot \mathbf{x}_i + w_0) \geq 1 \end{aligned}$$

- This we can solve! How?
  - It is a *quadratic programming* (QP) problem—a standard type of optimization problem for which many efficient packages are available.
  - Better yet, it's a convex (positive semidefinite) QP

Example





We have a solution, but no support vectors yet...

Lagrange multipliers for inequality constraints (revisited)

- Suppose we have the following optimization problem, called *primal*:

$$\begin{aligned} &\min_{\mathbf{w}} f(\mathbf{w}) \\ \text{such that } &g_i(\mathbf{w}) \quad \leq \quad 0, \; i = 1 \dots k \end{aligned}$$

- We define the *generalized Lagrangian*:

$$L(\mathbf{w}, \alpha) = f(\mathbf{w}) + \sum_{i=1}^{k} \alpha_i g_i(\mathbf{w}), \tag{1}$$

where $\alpha_i, \; i = 1 \dots k$ are the Lagrange multipliers.

A different optimization problem

- Consider $\mathcal{P}(\mathbf{w}) = \max_{\alpha:\alpha_i \geq 0} L(\mathbf{w}, \alpha)$
- Observe that the follow is true (see extra notes):

$$\mathcal{P}(\mathbf{w}) = \begin{cases} f(\mathbf{w}) & \text{if all constraints are satisfied} \\ +\infty & \text{otherwise} \end{cases}$$

- Hence, instead of computing $\min_{\mathbf{w}} f(\mathbf{w})$ subject to the original constraints, we can compute:

$$p^* = \min_{\mathbf{w}} \mathcal{P}(\mathbf{w}) = \min_{\mathbf{w}} \max_{\alpha:\alpha_i \geq 0} L(\mathbf{w}, \alpha)$$

Dual optimization problem

- Let $d^* = \max_{\alpha:\alpha_i \geq 0} \min_{\mathbf{w}} L(\mathbf{w}, \alpha)$ (max and min are reversed)
- We can show that $d^* \leq p^*$.
  - Let $p^* = L(w^p, \alpha^p)$
  - Let $d^* = L(w^d, \alpha^d)$
  - Then $d^* = L(w^d, \alpha^d) \leq L(w^p, \alpha^d) \leq L(w^p, \alpha^p) = p^*$.

Dual optimization problem

- If f, g_i are convex and the g_i can all be satisfied simultaneously for some $\mathbf{w}$, then we have equality: $d^* = p^* = L(\mathbf{w}^*, \alpha^*)$
- Moreover $\mathbf{w}^*, \alpha^*$ solve the primal and dual if and only if they satisfy the following conditions (called Karush-Kunh-Tucker):

$$\frac{\partial}{\partial w_i} L(\mathbf{w}^*, \alpha^*) = 0, \; i = 1 \ldots n \quad (2)$$

$$\alpha_i^* g_i(\mathbf{w}^*) = 0, \; i = 1 \ldots k \quad (3)$$

$$g_i(\mathbf{w}^*) \leq 0, \; i = 1 \ldots k \quad (4)$$

$$\alpha_i^* \geq 0, \; i = 1 \ldots k \quad (5)$$

Back to maximum margin perceptron

- We wanted to solve (rewritten slightly):

$$\begin{aligned} \text{min} \quad & \tfrac{1}{2}\|\mathbf{w}\|^2 \\ \text{w.r.t.} \quad & \mathbf{w}, w_0 \\ \text{s.t.} \quad & 1 - y_i(\mathbf{w} \cdot \mathbf{x}_i + w_0) \leq 0 \end{aligned}$$

- The Lagrangian is:

$$L(\mathbf{w}, w_0, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_i \alpha_i (1 - y_i(\mathbf{w} \cdot \mathbf{x}_i + w_0))$$

- The primal problem is: $\min_{\mathbf{w}, w_0} \max_{\alpha : \alpha_i \geq 0} L(\mathbf{w}, w_0, \alpha)$
- We will solve the dual problem: $\max_{\alpha : \alpha_i \geq 0} \min_{\mathbf{w}, w_0} L(\mathbf{w}, w_0, \alpha)$
- In this case, the optimal solutions coincide, because we have a quadratic objective and linear constraints (both of which are convex).

Solving the dual

- From KKT (2), the derivatives of $L(\mathbf{w}, w_0, \alpha)$ wrt $\mathbf{w}, w_0$ should be 0
- The condition on the derivative wrt w_0 gives $\sum_i \alpha_i y_i = 0$
- The condition on the derivative wrt $\mathbf{w}$ gives:

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x_i}$$

$\Rightarrow$ Just like for the perceptron with zero initial weights, the optimal solution for $\mathbf{w}$ is a linear combination of the $\mathbf{x}_i$, and likewise for w_0.

- The output is

$$h_{\mathbf{w},w_0}(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}) + w_0\right)$$

$\Rightarrow$ Output depends on weighted dot product of input vector with training examples

Solving the dual (II)

- By plugging these back into the expression for L, we get:

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

  with constraints: $\alpha_i \geq 0$ and $\sum_i \alpha_i y_i = 0$

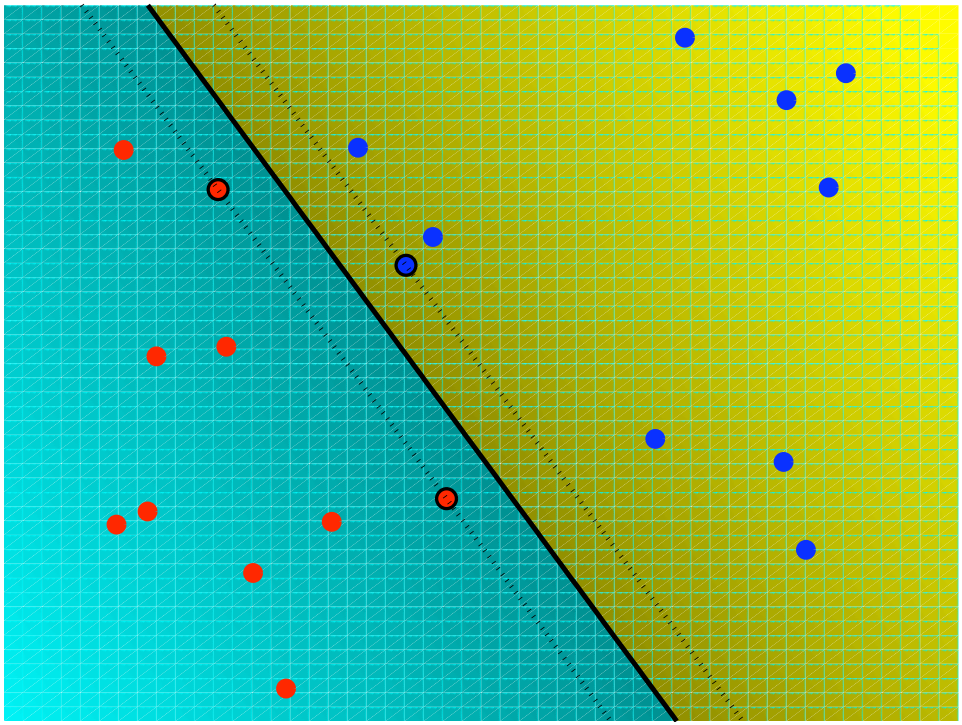
The support vectors

- Suppose we find optimal αs (e.g., using a standard QP package)
- The α_i will be > 0 only for the points for which $1 - y_i(\mathbf{w} \cdot \mathbf{x}_i + w_0) = 0$
- These are the points lying on the edge of the margin, and they are called *support vectors*, because they define the decision boundary
- The output of the classifier for query point $\mathbf{x}$ is computed as:

$$\text{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}) + w_0\right)$$

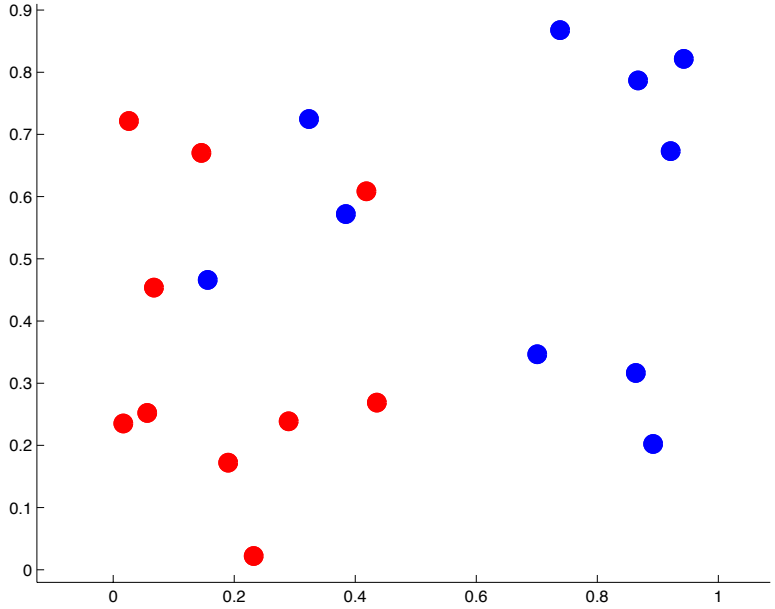
Hence, the output is determined by computing the *dot product of the point with the support vectors*!

Example



Support vectors are in bold

Non-linearly separable data



- A linear boundary might be too simple to capture the class structure.
- One way of getting a nonlinear decision boundary in the input space is to find a linear decision boundary in an expanded space (e.g., for polynomial regression.)
- Thus, $\mathbf{x}_i$ is replaced by $\phi(\mathbf{x}_i)$, where ϕ is called a *feature mapping*

Margin optimization in feature space

- Replacing $\mathbf{x}_i$ with $\phi(\mathbf{x}_i)$, the optimization problem to find $\mathbf{w}$ and w_0 becomes:

$$\begin{array}{rl} \text{min} & \|\mathbf{w}\|^2 \\ \text{w.r.t.} & \mathbf{w}, w_0 \\ \text{s.t.} & \mathbf{y}_i(\mathbf{w} \cdot \phi(\mathbf{x}_i) + w_0) \geq 1 \end{array}$$

- Dual form:

$$\begin{array}{rl} \text{max} & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \mathbf{y}_i \mathbf{y}_j \alpha_i \alpha_j \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \\ \text{w.r.t.} & \alpha_i \\ \text{s.t.} & 0 \leq \alpha_i \\ & \sum_{i=1}^m \alpha_i \mathbf{y}_i = 0 \end{array}$$

Feature space solution

- The optimal weights, in the expanded feature space, are $\mathbf{w} = \sum_{i=1}^{m} \alpha_i \mathbf{y}_i \phi(\mathbf{x}_i)$.

- Classification of an input $\mathbf{x}$ is given by:

$$h_{\mathbf{w},w_0}(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i \mathbf{y}_i \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}) + w_0\right)$$

$\Rightarrow$ Note that to solve the SVM optimization problem in dual form and to make a prediction, we only ever need to compute *dot-products of feature vectors*.

Kernel functions

- Whenever a learning algorithm (such as SVMs) can be written in terms of dot-products, it can be generalized to kernels.
- A *kernel* is any function $K : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ which corresponds to a dot product for some feature mapping ϕ:

$$K(\mathbf{x}_1, \mathbf{x}_2) = \phi(\mathbf{x}_1) \cdot \phi(\mathbf{x}_2) \text{ for some } \phi.$$

- Conversely, by choosing feature mapping ϕ, we implicitly choose a kernel function
- Recall that $\phi(\mathbf{x}_1) \cdot \phi(\mathbf{x}_2) = \cos \angle(\mathbf{x}_1, \mathbf{x}_2)$ where $\angle$ denotes the angle between the vectors, so a kernel function can be thought of as a notion of *similarity*.

The "kernel trick"

- If we work with the dual, we do not actually have to ever compute the feature mapping ϕ. We just have to compute the similarity K.
- That is, we can solve the dual for the α_i:

$$\begin{aligned} \text{max} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \mathbf{y}_i \mathbf{y}_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \text{w.r.t.} \quad & \alpha_i \\ \text{s.t.} \quad & 0 \leq \alpha_i \\ & \sum_{i=1}^m \alpha_i \mathbf{y}_i = 0 \end{aligned}$$

- The class of a new input $\mathbf{x}$ is computed as:

$$h_{\mathbf{w},w_0}(\mathbf{x}) = \text{sign}\left((\sum_{i=1}^m \alpha_i y_i \phi(\mathbf{x}_i)) \cdot \phi(\mathbf{x}) + w_0 \right) = \text{sign}\left(\sum_{i=1}^m \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + w_0 \right)$$

- Often, $K(\cdot, \cdot)$ can be evaluated in $O(n)$ time—a big savings!

Regularization with SVMs

- Kernels are a powerful tool for allowing non-linear, complex functions
- But now the number of parameters can be as high as the number of instances!
- With a very specific, non-linear kernel, each data point may be in its own partition
- With linear and logistic regression, we used regularization to avoid overfitting
- We need a method for allowing regularization with SVMs as well.

Soft margin classifiers

- Recall that in the linearly separable case, we compute the solution to the following optimization problem:

$$\begin{aligned} \text{min} \quad & \tfrac{1}{2}\|\mathbf{w}\|^2 \\ \text{w.r.t.} \quad & \mathbf{w}, w_0 \\ \text{s.t.} \quad & y_i(\mathbf{w} \cdot \mathbf{x}_i + w_0) \geq 1 \end{aligned}$$

- If we want to allow misclassifications, we can relax the constraints to:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + w_0) \geq 1 - \xi_i$$

- If $\xi_i \in (0, 1)$, the data point is within the margin
- If $\xi_i \geq 1$, then the data point is misclassified
- We define the *soft error* as $\sum_i \xi_i$
- We will have to change the criterion to reflect the soft errors

New problem formulation with soft errors

- Instead of:

$$\begin{aligned} \text{min} \quad & \tfrac{1}{2}\|\mathbf{w}\|^2 \\ \text{w.r.t.} \quad & \mathbf{w}, w_0 \\ \text{s.t.} \quad & y_i(\mathbf{w}\cdot\mathbf{x}_i + w_0) \geq 1 \end{aligned}$$

  we want to solve:

$$\begin{aligned} \text{min} \quad & \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \\ \text{w.r.t.} \quad & \mathbf{w}, w_0, \xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}\cdot\mathbf{x}_i + w_0) \geq 1 - \xi_i,\ \xi_i \geq 0 \end{aligned}$$

- Note that soft errors include points that are misclassified, as well as points within the margin
- There is a linear penalty for both categories
- The choice of the *constant C controls overfitting*

A built-in overfitting knob

$$\begin{array}{rl} \text{min} & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \\ \text{w.r.t.} & \mathbf{w}, w_0, \xi_i \\ \text{s.t.} & y_i(\mathbf{w}\cdot\mathbf{x}_i + w_0) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{array}$$

- If C is 0, there is no penalty for soft errors, so the focus is on maximizing the margin, even if this means more mistakes
- If C is very large, the emphasis on the soft errors will cause decreasing the margin, if this helps to classify more examples correctly.
- Internal cross-validation is a good way to choose C appropriately

Lagrangian for the new problem

- Like before, we can write a Lagrangian for the problem and then use the dual formulation to find the optimal parameters:

$$\begin{aligned} L(\mathbf{w}, w_0, \alpha, \xi, \mu) &= \frac{1}{2}||\mathbf{w}||^2 + C\sum_i \xi_i \\ &+ \sum_i \alpha_i \left(1 - \xi_i - y_i(\mathbf{w}_i \cdot \mathbf{x}_i + w_0)\right) + \sum_i \mu_i \xi_i \end{aligned}$$

- All the previously described machinery can be used to solve this problem
- Note that in addition to α_i we have coefficients μ_i, which ensure that the errors are positive, but do not participate in the decision boundary

Soft margin optimization with kernels

- Replacing $\mathbf{x}_i$ with $\phi(\mathbf{x}_i)$, the optimization problem to find $\mathbf{w}$ and w_0 becomes:

$$\begin{aligned} \text{min} \quad & \|\mathbf{w}\|^2 + C\sum_i \zeta_i \\ \text{w.r.t.} \quad & \mathbf{w}, w_0, \zeta_i \\ \text{s.t.} \quad & \mathbf{y}_i(\mathbf{w}\cdot\phi(\mathbf{x}_i) + w_0) \geq (1-\zeta_i) \\ & \zeta_i \geq 0 \end{aligned}$$

- Dual form and solution have similar forms to what we described last time, but in terms of kernels

Getting SVMs to work in practice

- Two important choices:
  - Kernel (and kernel parameters)
  - Regularization parameter C
- The parameters may interact!

  E.g. for Gaussian kernel, the larger the width of the kernel, the more biased the classifier, so low C is better
- Together, these control overfitting: always do an internal parameter search, using a validation set!
- Overfitting symptoms:
  - Low margin
  - Large fraction of instances are support vectors

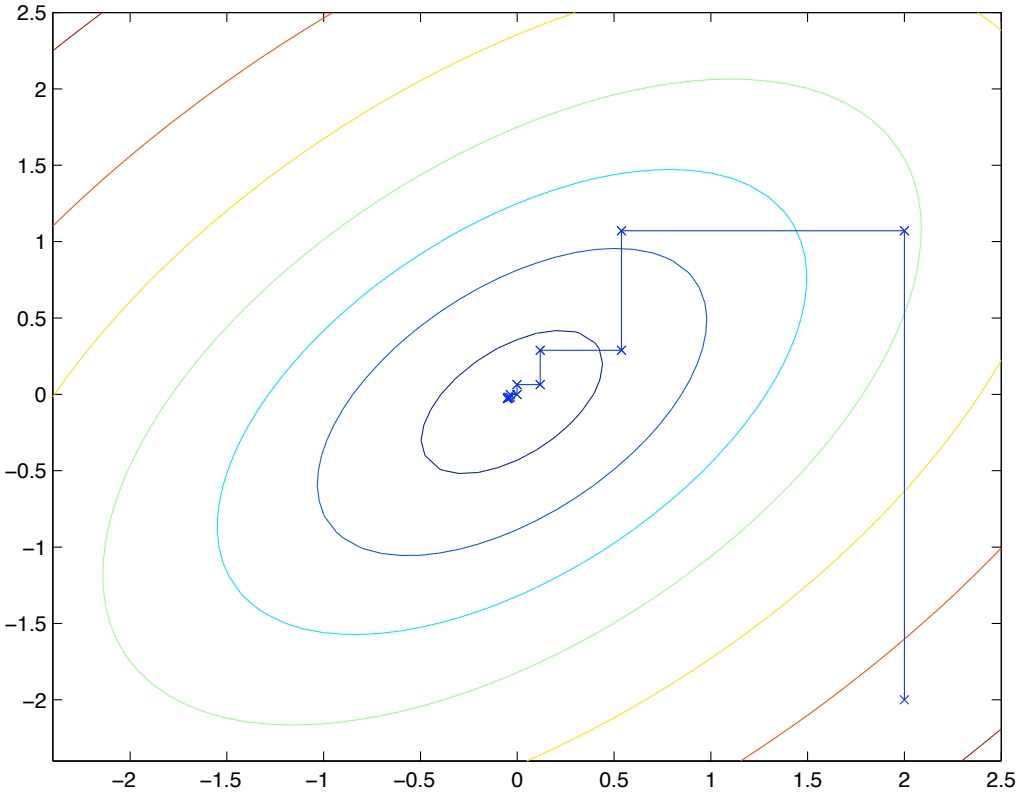
Solving the quadratic optimization problem

- Many approaches exist
- Because we have constraints, gradient descent does not apply directly (the optimum might be outside of the feasible region)
- Platt's algorithm is the fastest current approach, based on *coordinate ascent*

Coordinate ascent

- Suppose you want to find the maximum of some function $F(\alpha_1, \ldots \alpha_n)$
- Coordinate ascent optimizes the function by repeatedly picking an α_i and optimizing it, while all other parameters are fixed
- There are different ways of looping through the parameters:
  - Round-robin
  - Repeatedly pick a parameter at random
  - Choose next the variable expected to make the largest improvement
  - ...

Example



Our optimization problem (dual form)

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

with constraints: $0 \leq \alpha_i \leq C$ and $\sum_i \alpha_i y_i = 0$

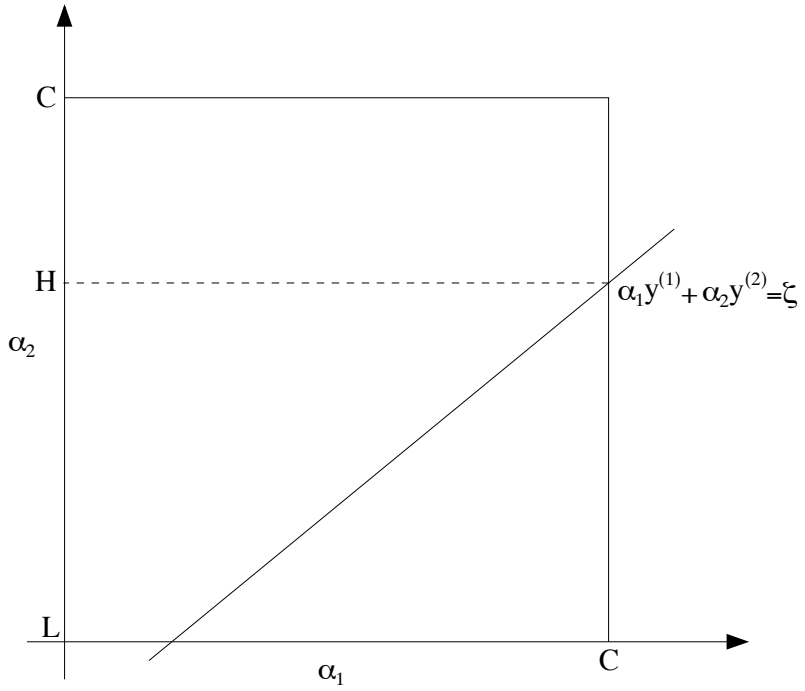
- Suppose we want to optimize for α_1 while $\alpha_2, \ldots \alpha_n$ are fixed
- We cannot do it because α_1 will be completely determined by the last constraint: $\alpha_1 = -y_1 \sum_{i=2}^{m} \alpha_i y_i$
- Instead, we have to optimize *pairs of parameters* α_i, α_j together

SMO

- Suppose that we want to optimize α_1 and α_2 together, while all other parameters are fixed.
- We know that $y_1\alpha_1 + y_2\alpha_2 = -\sum_{i=1}^{m} y_i\alpha_i = \xi$, where ξ is a constant
- So $\alpha_1 = y_1(\xi - y_2\alpha_2)$ (because y_1 is either $+1$ or -1 so $y_1^2 = 1$)
- This defines a line, and any pair α_1, α_2 which is a solution has to be on the line

SMO (II)

- We also know that $0 \leq \alpha_1 \leq C$ and $0 \leq \alpha_2 \leq C$, so the solution has to be on the line segment inside the rectangle below



SMO(III)

- By plugging α_1 back in the optimization criterion, we obtain a quadratic function of α_2, whose optimum we can find exactly
- If the optimum is inside the rectangle, we take it.
- If not, we pick the closest intersection point of the line and the rectangle
- This procedure is very fast because all these are simple computations.

Interpretability

- SVMs are not very intuitive, but typically are more interpretable than neural nets, if you look at the machine and the misclassifications
- E.g. Ovarian cancer data (Haussler) - 31 tissue samples of 3 classes, misclassified examples wrongly labelled
- But no biological plausibility!
- Hard to interpret if the percentage of instances that are recruited as support vectors is high

Complexity

- Quadratic programming is expensive in the number of training examples
- Platt's SMO algorithm is quite fast though, and other fancy optimization approaches are available
- Best packages can handle $50,000+$ instances, but not more than $100,000$
- On the other hand, number of attributes can be very high (strength compared to neural nets)
- Evaluating a SVM is *slow if there are a lot of support vectors*.
- Dictionary methods attempt to select a subset of the data on which to train.