Machine Learning and Intelligent Systems

Kernel Machines

Maria A. Zuluaga

Dec 6, 2022

EURECOM - Data Science Department

Table of contents

Kernel Machines

Kernel Machines

- We have introduced kernels and showed that they can be a very powerful tool
- The remaining question is how can we use them?
- Given the linear models that we have seen, how can we integrate the use of kernels in them?

Kernelizing an algorithm

To kernelize an algorithm, three steps are required:

Kernelizing an algorithm

To kernelize an algorithm, three steps are required:

1. Demonstrate that the solution lies in the span of the training data, i.e.

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i \tag{1}$$

   for some α_i.

Kernelizing an algorithm

To kernelize an algorithm, three steps are required:

1. Demonstrate that the solution lies in the span of the training data, i.e.

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i \tag{1}$$

   for some α_i.

2. Rewrite the algorithm and the model so that the train/text inputs are only accessed via inner-products, i.e.

$$\mathbf{x}_i^T \mathbf{x}_j$$

Kernelizing an algorithm

To kernelize an algorithm, three steps are required:

1. Demonstrate that the solution lies in the span of the training data, i.e.

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i \tag{1}$$

   for some α_i.

2. Rewrite the algorithm and the model so that the train/text inputs are only accessed via inner-products, i.e.

$$\mathbf{x}_i^T \mathbf{x}_j$$

3. Define a kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$

Example 1: Kernel Linear Regression (OLS)

Recap

- The OLS solution minimizes the quadratic loss:

$$\arg\min_{\mathbf{w}} \sum_{i=1}^{N} (\mathbf{w}^T \mathbf{x}_i - y_i)^2$$

- It was closed form solution

$$\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X} \mathbf{y}$$

- The prediction of an unseen point is done via:

$$h(\mathbf{x}^*) = \hat{\mathbf{w}}^T \mathbf{x}^*$$

Step 1: Prove the solution is a linear combination of the inputs

- Let us express $\mathbf{w}$ as a linear combination of the inputs

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i$$
$$= \mathbf{X}^T \vec{\alpha}$$

Step 1: Prove the solution is a linear combination of the inputs

- Let us express $\mathbf{w}$ as a linear combination of the inputs

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i$$
$$= \mathbf{X}^T \vec{\alpha}$$

- Since the squared loss is a convex function, last lecture we demonstrated that such a solution exists
- It is obtained by applying gradient descent and initializing $\vec{\alpha} = 0$

Step 2: Rewrite in terms of inner products

- Kernelization of the prediction step is trivial:

$$
\begin{aligned}
h(\mathbf{x}^*) &= \hat{\mathbf{w}}^T \mathbf{x}^* \\
&= \sum_{i=1}^{N} \alpha_i \mathbf{x}_i^T \mathbf{x}^*
\end{aligned}
$$

- Kernelization is trivial as it requires to replace inner products by $k(\cdot, \cdot)$

$$
h(\mathbf{x}^*) = \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_i, \mathbf{x}^*)
$$

Closed form solution of $\vec{\alpha}$

As with $\hat{\mathbf{w}}$, the kernelized version of OLS allows for a closed form solution for $\vec{\alpha}$

Closed form solution of $\vec{\alpha}$

As with $\hat{\mathbf{w}}$, the kernelized version of OLS allows for a closed form solution for $\vec{\alpha}$

Theorem:

Kernel Ordinary Least Squares has the solution:

$$\vec{\alpha} = \mathbf{K}^{-1}\mathbf{y}$$

Closed form solution of $\vec{\alpha}$

As with $\hat{\mathbf{w}}$, the kernelized version of OLS allows for a closed form solution for $\vec{\alpha}$

Theorem:

Kernel Ordinary Least Squares has the solution:

$$\vec{\alpha} = \mathbf{K}^{-1}\mathbf{y}$$

Proof:

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Closed form solution of $\vec{\alpha}$

As with $\hat{\mathbf{w}}$, the kernelized version of OLS allows for a closed form solution for $\vec{\alpha}$

Theorem:

Kernel Ordinary Least Squares has the solution:

$$\vec{\alpha} = \mathbf{K}^{-1}\mathbf{y}$$

Proof:

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \rightarrow \mathbf{X}^T\vec{\alpha}$$

Closed form solution of $\vec{\alpha}$

As with $\hat{\mathbf{w}}$, the kernelized version of OLS allows for a closed form solution for $\vec{\alpha}$

Theorem:

Kernel Ordinary Least Squares has the solution:

$$\vec{\alpha} = \mathbf{K}^{-1}\mathbf{y}$$

Proof:

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \rightarrow \mathbf{X}^T\vec{\alpha}$$

Exercise: Do the same exercise to obtain kernel ridge regression

Example 2: Kernel Support Vector Machines

Recap: Hard SVM

The solution of the hard SVM required solving a constrained optimization problem:

$$\begin{aligned} \underset{\mathbf{w}, w_0}{\arg\min} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{subject to} \quad & \forall i \ y_i(\mathbf{w}^T \mathbf{x}_i + w_0) \geq 1 \end{aligned}$$

Prediction of a new point at testing is:

$$h(\mathbf{x}^*) = \text{sign}(\hat{\mathbf{w}}^T \mathbf{x}^* + \hat{w}_0)$$

Formulating kernel SVM requires some manipulations

Dual Form of an Optimization Problem

- An optimization problem has a **dual form** if the function to be optimized and the constraints are strictly convex
- If this is the case, the dual form is also a solution of the primal form of the optimization problem

Dual Form of an Optimization Problem

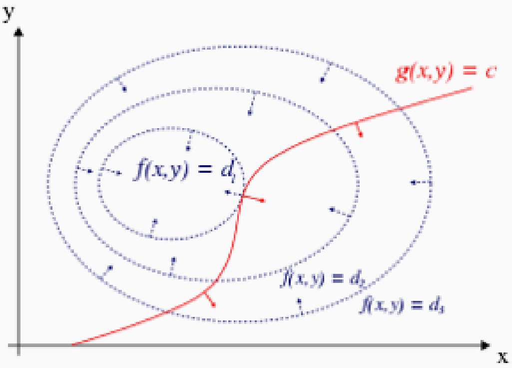
- An optimization problem has a **dual form** if the function to be optimized and the constraints are strictly convex
- If this is the case, the dual form is also a solution of the primal form of the optimization problem
- Usually the term dual problem refers to the Lagrangian dual problem but other dual problems are used

Dual Form of an Optimization Problem

- An optimization problem has a **dual form** if the function to be optimized and the constraints are strictly convex
- If this is the case, the dual form is also a solution of the primal form of the optimization problem
- Usually the term dual problem refers to the Lagrangian dual problem but other dual problems are used
- The Lagrangian dual problem is obtained by forming the Lagrangian of a minimization problem by using non-negative Lagrange multipliers to add the constraints to the objective function, and then solving for the primal variable values that minimize the original objective function

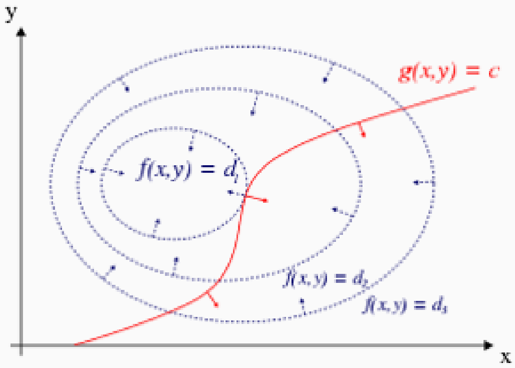
Solving the Constrained Optimization Problem

- **Problem formulation:** We want to optimizer $f(\cdot)$ subject to a constraint $g(\cdot) = c$



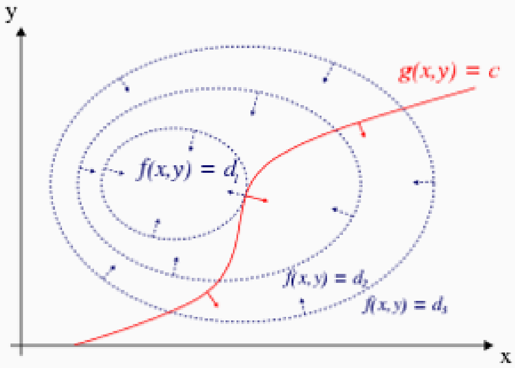
Solving the Constrained Optimization Problem

- **Problem formulation:** We want to optimizer $f(\cdot)$ subject to a constraint $g(\cdot) = c$
- Optimizing f s.t. $g(\cdot) = c$ means to find the level curve of f with **maximum** d_i value intersecting the constraint curve
- At this point, the two curves are tangent



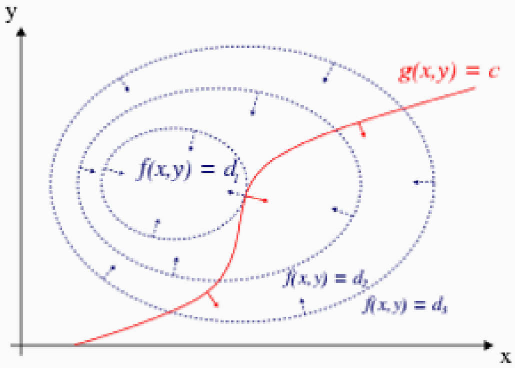
Solving the Constrained Optimization Problem

- **Problem formulation:** We want to optimizer $f(\cdot)$ subject to a constraint $g(\cdot) = c$
- Optimizing f s.t. $g(\cdot) = c$ means to find the level curve of f with **maximum** d_i value intersecting the constraint curve
- At this point, the two curves are tangent
- **Fact:** Two curves have a common perpendicular line if they are tangent at that point



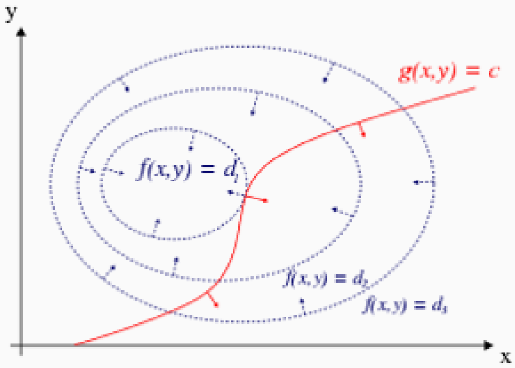
Solving the Constrained Optimization Problem

- **Problem formulation:** We want to optimizer $f(\cdot)$ subject to a constraint $g(\cdot) = c$
- Optimizing f s.t. $g(\cdot) = c$ means to find the level curve of f with **maximum** d_i value intersecting the constraint curve
- At this point, the two curves are tangent
- **Fact:** Two curves have a common perpendicular line if they are tangent at that point
- ∇f is perpendicular to its level curves
- ∇g is perpendicular to the constraint curve



Solving the Constrained Optimization Problem

- **Problem formulation:** We want to optimizer $f(\cdot)$ subject to a constraint $g(\cdot) = c$
- Optimizing f s.t. $g(\cdot) = c$ means to find the level curve of f with **maximum** d_i value intersecting the constraint curve
- At this point, the two curves are tangent
- **Fact:** Two curves have a common perpendicular line if they are tangent at that point
- ∇f is perpendicular to its level curves
- ∇g is perpendicular to the constraint curve



Idea:

Find points where $\nabla f + \lambda \nabla g = 0$

Lagrange Method

First case: One equality constraint, $f(\mathbf{x})$ s.t. $g(\mathbf{x}) = c$

1. Define the Lagrangian function:

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$$

   with $\lambda \neq 0$ denoted the Lagrange multiplier. This is the new function to maximize

2. Satisfy the constrained stationarity condition, i.e. $\nabla f + \lambda \nabla g = 0$ through

$$\nabla_{\mathbf{x}} L = 0$$

3. Satisfy the constraint equation $g(\mathbf{x}) = 0$ through

$$\frac{\partial L}{\partial \lambda} = 0$$

One equality constraint: Example

Find a stationary point of the function $f(\mathbf{x}) = 1 - x_1^2 - x_2^2$ s.t. $x_1 + x_2 = 1$

1. Identify f and g

$$f(\mathbf{x}) = 1 - x_1^2 - x_2^2, \qquad g(\mathbf{x}) = x_1 + x_2 - 1 = 0$$

2. Define the Lagrangian function

$$L(\mathbf{x}, \lambda) = 1 - x_1^2 - x_2^2 + \lambda(x_1 + x_2 - 1)$$

3. Express $\nabla_{\mathbf{x}} L = 0$

$$\frac{\partial L}{\partial x_1} = -2x_1 + \lambda = 0 \qquad \frac{\partial L}{\partial x_2} = -2x_2 + \lambda = 0$$

4. Express $\partial L / \partial \lambda = 0$

$$\frac{\partial L}{\partial \lambda} = x_1 + x_2 - 1 = 0$$

One equality constraint: Example

We obtain a 3×3 system of equations and unknowns that can be solved through simple arithmetic:

$$\frac{\partial L}{\partial x_1} = -2x_1 + \lambda = 0$$
$$\frac{\partial L}{\partial x_2} = -2x_2 + \lambda = 0$$
$$\frac{\partial L}{\partial \lambda} = x_1 + x_2 - 1 = 0$$

with the solution $(\hat{x_1}, \hat{x_2}) = (1/2, 1/2)$ and $\lambda = 1$

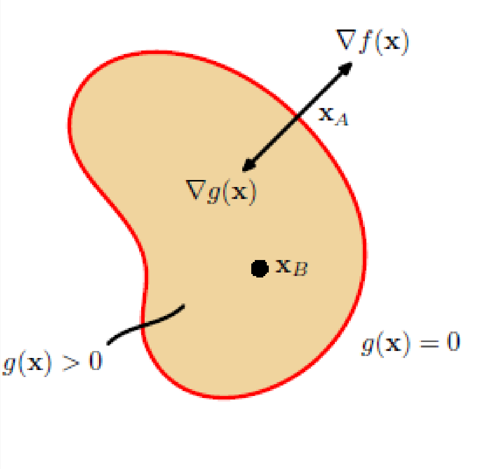
Let's now move to the case where the constraint is an inequality

Lagrange Method

Second case: One inequality constraint, $f(\mathbf{x})$ s.t. $g(\mathbf{x}) \geq 0$

There are two kind of solutions possible:

- **Inactive constraint:** The stationary point lies in the region where $g(\mathbf{x}) > 0$



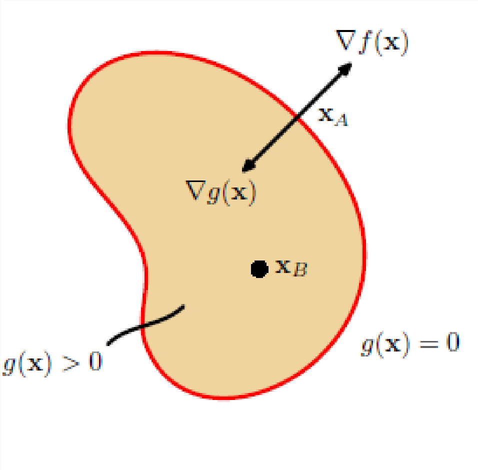
Source: Fig E.3 in PRML

Lagrange Method

Second case: One inequality constraint, $f(\mathbf{x})$ s.t. $g(\mathbf{x}) \geq 0$

There are two kind of solutions possible:

- **Inactive constraint:** The stationary point lies in the region where $g(\mathbf{x}) > 0$
- **Active constraint:** The stationary point lies in the boundary $g(\mathbf{x}) = 0$

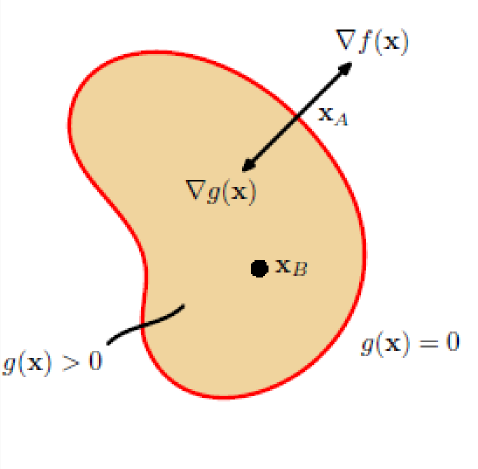


Source: Fig E.3 in PRML

Lagrange Method: Inactive Constraint

Inactive constraint:

- The stationarity condition, $\nabla f + \lambda \nabla g = 0$, plays no role

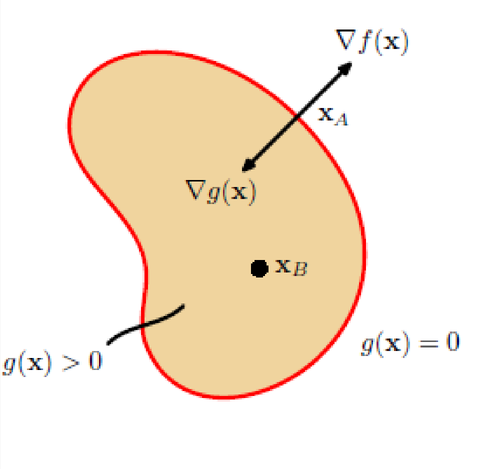


Source: Fig E.3 in PRML

Lagrange Method: Inactive Constraint

Inactive constraint:

- The stationarity condition, $\nabla f + \lambda \nabla g = 0$, plays no role
- It can be expressed as $\nabla_{\mathbf{x}} f = 0$

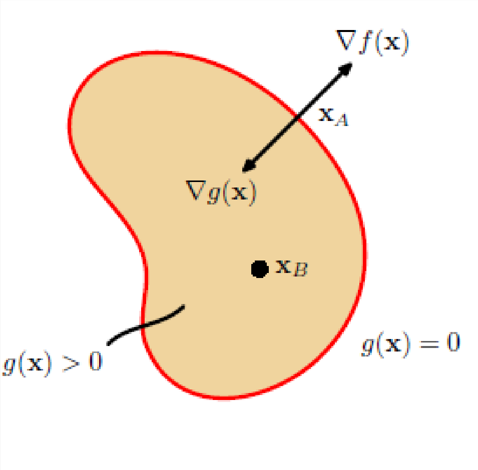


Source: Fig E.3 in PRML

Lagrange Method: Inactive Constraint

Inactive constraint:

- The stationarity condition, $\nabla f + \lambda \nabla g = 0$, plays no role
- It can be expressed as $\nabla_{\mathbf{x}} f = 0$
- This corresponds to a stationary point of the Lagrange function with $\lambda = 0$

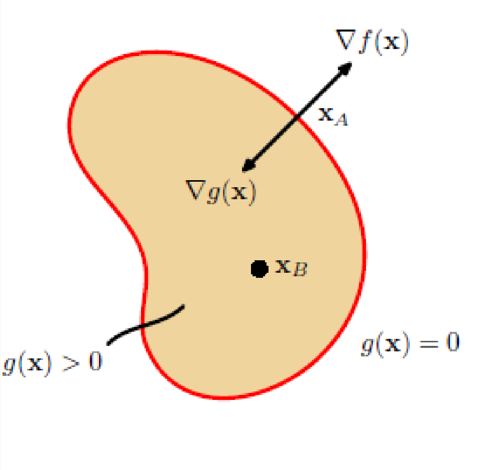


Source: Fig E.3 in PRML

Lagrange Method: Active Constraint

Active constraint:

- Analogous to the equality constraint, $\lambda \neq 0$

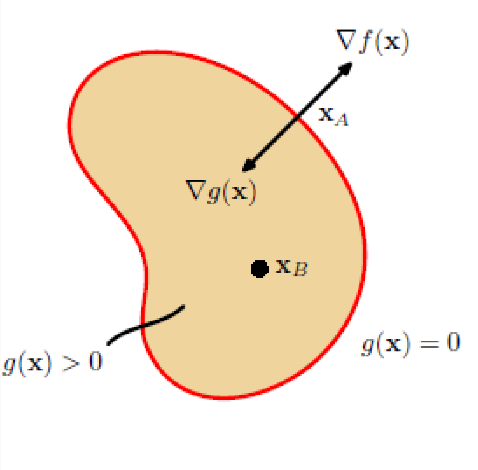


Source: Fig E.3 in PRML

Lagrange Method: Active Constraint

Active constraint:

- Analogous to the equality constraint, $\lambda \neq 0$
- The sign of λ is important

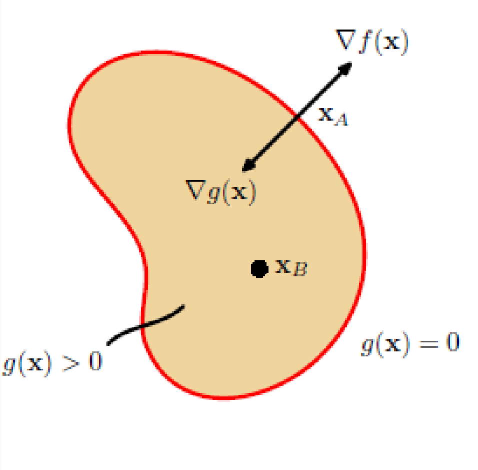


Source: Fig E.3 in PRML

Lagrange Method: Active Constraint

Active constraint:

- Analogous to the equality constraint, $\lambda \neq 0$
- The sign of λ is important
- f will only be at a maximum if its gradient is oriented away from the region $g > 0$

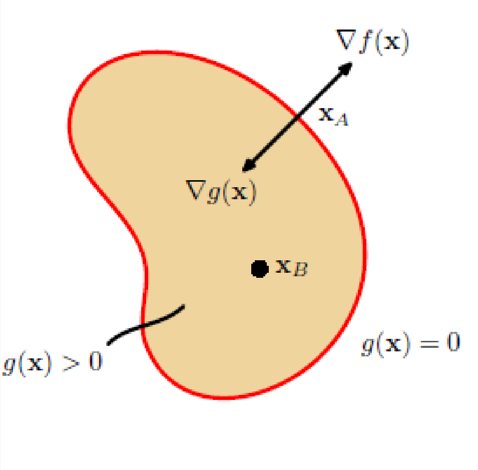


Source: Fig E.3 in PRML

Lagrange Method: Active Constraint

Active constraint:

- Analogous to the equality constraint, $\lambda \neq 0$
- The sign of λ is important
- f will only be at a maximum if its gradient is oriented away from the region $g > 0$
- This implies $\nabla f = -\lambda \nabla g$, $\lambda > 0$



Source: Fig E.3 in PRML

Lagrange Method

- For both cases (active and inactive) it holds that $\lambda g(\mathbf{x}) = 0$
- Therefore, the solution of maximizing $f(\mathbf{x})$ s.t. $g(\mathbf{x}) \geq 0$ is obtained by maximizing the Lagrange function with respect to $\mathbf{x}$, λ subject to the constraints:

$$g(\mathbf{x}) \geq 0 \tag{2}$$

$$\lambda \geq 0 \tag{3}$$

$$\lambda g(\mathbf{x}) = 0 \tag{4}$$

- These are known as the Karush-Kuhn-Tucker (KKT) conditions

Lagrange Method

- For both cases (active and inactive) it holds that $\lambda g(\mathbf{x}) = 0$
- Therefore, the solution of maximizing $f(\mathbf{x})$ s.t. $g(\mathbf{x}) \geq 0$ is obtained by maximizing the Lagrange function with respect to $\mathbf{x}$, λ subject to the constraints:

$$g(\mathbf{x}) \geq 0 \tag{2}$$

$$\lambda \geq 0 \tag{3}$$

$$\lambda g(\mathbf{x}) = 0 \tag{4}$$

- These are known as the Karush-Kuhn-Tucker (KKT) conditions

Note 1: If the task is to minimize subject to $g \geq 0$, the Lagrangian function becomes:

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda g(\mathbf{x})$$

Note 2: Multiple equalities or inequalities are a trivial extension: one constraint, one set of Lagrange multipliers

Back to SVM

The solution of the hard SVM required solving a constrained optimization problem:

$$\begin{aligned}
\arg\min_{\mathbf{w}, w_0} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\
\text{subject to} \quad & \forall i \ y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1
\end{aligned}$$

Let's use the Lagrange method!

Lagrange Method for SVMs

1. Identify f and g

Lagrange Method for SVMs

1. Identify f and g

$$f(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 \qquad g_i(\mathbf{w}) = y_i(\mathbf{w}^T\mathbf{x}_i + w_0) - 1$$

Lagrange Method for SVMs

1. Identify f and g

$$f(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 \qquad g_i(\mathbf{w}) = y_i(\mathbf{w}^T\mathbf{x}_i + w_0) - 1$$

2. Express the Lagrangian function. Idea: Let's replace λ by α

Lagrange Method for SVMs

1. Identify f and g

$$f(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 \qquad g_i(\mathbf{w}) = y_i(\mathbf{w}^T\mathbf{x}_i + w_0) - 1$$

2. Express the Lagrangian function. Idea: Let's replace λ by α

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

Lagrange Method for SVMs

1. Identify f and g

$$f(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 \qquad g_i(\mathbf{w}) = y_i(\mathbf{w}^T\mathbf{x}_i + w_0) - 1$$

2. Express the Lagrangian function. Idea: Let's replace λ by α

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

3. Express $\nabla_{\mathbf{w}} L$, $\partial L/\partial w_0$

Lagrange Method for SVMs

1. Identify f and g

$$f(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 \qquad g_i(\mathbf{w}) = y_i(\mathbf{w}^T\mathbf{x}_i + w_0) - 1$$

2. Express the Lagrangian function. Idea: Let's replace λ by α

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

3. Express $\nabla_{\mathbf{w}} L$, $\partial L / \partial w_0$

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i = 0, \qquad \frac{\partial L}{\partial w_0} = -\sum_{i=1}^{N} \alpha_i y_i = 0$$

Lagrange Method for SVMs

1. Identify f and g

$$f(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 \qquad g_i(\mathbf{w}) = y_i(\mathbf{w}^T\mathbf{x}_i + w_0) - 1$$

2. Express the Lagrangian function. Idea: Let's replace λ by α

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

3. Express $\nabla_{\mathbf{w}} L$, $\partial L/\partial w_0$

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i = 0, \qquad \frac{\partial L}{\partial w_0} = -\sum_{i=1}^{N} \alpha_i y_i = 0$$

From this it follows that:

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i, \qquad \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{5}$$

The Lagrangian Function

Let's replace the terms obtained in Eq. 5 in the Lagrangian function:

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

Let's focus first on the first term:

$$\begin{aligned}
\|\mathbf{w}\|^2 &= \mathbf{w}^T \mathbf{w} \\
&= \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i^T \cdot \sum_{j=1}^{N} \alpha_j y_j \mathbf{x}_j \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j
\end{aligned}$$

The Lagrangian Function

The Lagrangian function now becomes:

$$L(\mathbf{w}, w_0, \alpha) = \frac{1}{2}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j\right) - \sum_{i=1}^{N}\alpha_i(y_i(\mathbf{w}^T\mathbf{x}_i + w_0) - 1)$$

The Lagrangian Function

The Lagrangian function now becomes:

$$L(\mathbf{w}, w_0, \alpha) = \frac{1}{2}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j\right) - \sum_{i=1}^{N}\alpha_i(y_i(\mathbf{w}^T\mathbf{x}_i + w_0) - 1)$$

Now let's do the second term. We split it and then replace for $\mathbf{w}$:

$$\begin{aligned}
\sum_{i=1}^{N}\alpha_i(y_i(\mathbf{w}^T\mathbf{x}_i + w_0) - 1) &= \sum_{i=1}^{N}\alpha_i y_i \mathbf{w}^T\mathbf{x}_i + \sum_{i=1}^{N}\underbrace{\alpha_i y_i}_{=0} w_0 - \sum_{i=1}^{N}\alpha_i \\
&= \sum_{i=1}^{N}\alpha_i y_i\left(\sum_{j=1}^{N}\alpha_j y_j \mathbf{x}_j^T\right)\mathbf{x}_i - \sum_{i=1}^{N}\alpha_i \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j - \sum_{i=1}^{N}\alpha_i
\end{aligned}$$

Dual Representation of the Maximum Margin Problem

Replacing the second term in the Lagrangian we obtain:

$$
\begin{aligned}
L(\boldsymbol{\alpha}) &= \frac{1}{2}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j\right) - \left(\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j - \sum_{i=1}^{N}\alpha_i\right) \\
&= \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j\right)
\end{aligned}
$$

This is the dual representation of the hard margin SVM optimization problem.

Dual Representation of the Maximum Margin Problem

Replacing the second term in the Lagrangian we obtain:

$$L(\boldsymbol{\alpha}) = \frac{1}{2}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j\right) - \left(\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j - \sum_{i=1}^{N}\alpha_i\right)$$

$$= \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j\right)$$

This is the dual representation of the hard margin SVM optimization problem.

The optimization of the dual problem becomes:

$$\begin{aligned} &\arg\max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j\right) \\ &\text{subject to} \quad \forall i\ \alpha_i \geq 0, \qquad \sum_{i=1}^{N}\alpha_i y_i = 0 \end{aligned} \tag{6}$$

Kernel SVM: Where are we?

- We showed that the training can be expressed in terms of inner products without formally expressing $\mathbf{w}$

Kernel SVM: Where are we?

- We showed that the training can be expressed in terms of inner products without formally expressing $\mathbf{w}$
- We showed that $\mathbf{w}$ can be expressed as a linear combination of the inputs

Kernel SVM: Where are we?

- We showed that the training can be expressed in terms of inner products without formally expressing $\mathbf{w}$
- We showed that $\mathbf{w}$ can be expressed as a linear combination of the inputs
- Since the original problem is convex, it is possible to train an SVM using gradient descent for a given set of $\boldsymbol{\alpha}$

Kernel SVM: Where are we?

- We showed that the training can be expressed in terms of inner products without formally expressing $\mathbf{w}$
- We showed that $\mathbf{w}$ can be expressed as a linear combination of the inputs
- Since the original problem is convex, it is possible to train an SVM using gradient descent for a given set of $\boldsymbol{\alpha}$
- The dual representation is a quadratic problem can be solved through standard solvers

Kernel SVM: Where are we?

- We showed that the training can be expressed in terms of inner products without formally expressing $\mathbf{w}$
- We showed that $\mathbf{w}$ can be expressed as a linear combination of the inputs
- Since the original problem is convex, it is possible to train an SVM using gradient descent for a given set of $\boldsymbol{\alpha}$
- The dual representation is a quadratic problem can be solved through standard solvers
- What we are missing?

Kernel SVM: Where are we?

- We showed that the training can be expressed in terms of inner products without formally expressing $\mathbf{w}$
- We showed that $\mathbf{w}$ can be expressed as a linear combination of the inputs
- Since the original problem is convex, it is possible to train an SVM using gradient descent for a given set of $\boldsymbol{\alpha}$
- The dual representation is a quadratic problem can be solved through standard solvers
- What we are missing?
    - w_0 is not estimated at training
    - How to make a prediction

Estimating w_0

We can use the third KKT condition to have an estimate of w_0:

$$\alpha g(\mathbf{x}) = 0$$

Estimating w_0

We can use the third KKT condition to have an estimate of w_0:

$$\alpha g(\mathbf{x}) = 0$$

Replacing what we identified as g:

$$\hat{\alpha}_i \left(y_i(\hat{\mathbf{w}}\mathbf{x}_i + \hat{w}_0) - 1 \right) = 0$$

$$\hat{\alpha}_i \left(y_i \left(\sum_{j=1}^{N} \alpha_j y_j \mathbf{x}_j^T \mathbf{x}_i + \hat{w}_0 \right) - 1 \right) = 0$$

Estimating w_0

We can use the third KKT condition to have an estimate of w_0:

$$\alpha g(\mathbf{x}) = 0$$

Replacing what we identified as g:

$$\hat{\alpha}_i \left(y_i(\hat{\mathbf{w}}\mathbf{x}_i + \hat{w}_0) - 1 \right) = 0$$

$$\hat{\alpha}_i \left(y_i \left(\sum_{j=1}^{N} \alpha_j y_j \mathbf{x}_j^T \mathbf{x}_i + \hat{w}_0 \right) - 1 \right) = 0$$

Let us recall that if a given $\mathbf{x}$ is a support vector then $y_i(\hat{\mathbf{w}}\mathbf{x}_i + \hat{w}_0) = 1$, so

$$\hat{w}_0 = y_i - \left(\sum_{j=1}^{N} \alpha_j y_j \mathbf{x}_j^T \mathbf{x}_i \right) = y_i - \left(\sum_{j=1}^{N} \alpha_j y_j k(\mathbf{x}_j, \mathbf{x}_i) \right) \tag{7}$$

In practice it is better to obtain $\hat{w}_0$ by averaging over all i's that are a support vector.

Prediction

The prediction of a new point is straightforward:

$$h(\mathbf{x}^*) = \text{sign}(\hat{\mathbf{w}}^T \mathbf{x}^* + \hat{w}_0) \tag{8}$$

It accounts to replacing Eqs. 5 and 7 in the term above.

Prediction

The prediction of a new point is straightforward:

$$h(\mathbf{x}^*) = \text{sign}(\hat{\mathbf{w}}^T \mathbf{x}^* + \hat{w}_0) \tag{8}$$

It accounts to replacing Eqs. 5 and 7 in the term above.

$$h(\mathbf{x}^*) = \text{sign}\left(\sum_{j=1}^{N} \alpha_j y_j k(\mathbf{x}_j, \mathbf{x}^*) + \left(y_i - \left(\sum_{j=1}^{N} \alpha_j y_j k(\mathbf{x}_j, \mathbf{x}_i)\right)\right)\right)$$

A final word on support vectors

- Given the KKT conditions, every point in the dataset satisfies either $\alpha_i = 0$ or $g(\cdot) = 1$.
- Hence any point where $\alpha_i = 0$ is not considered in the predictions.
- These points are the **support vectors**

Wrap-up

Wrap-up

- We presented the necessary steps to transform a given method to handle kernels
- We used the ordinary least squares as a first example
- We reviewed the Lagrange method and used it to formulate the primal and dual hard SVM optimization problems
- We used Lagrange multipliers to formulate the Kernel SVM

Key Concepts

- Kernel OLS
- Lagrange function
- Lagrange multipliers
- Dual representation
- Kernel SVM

References

Further Reading and Useful Material

Source	Notes
Pattern Recognition and Machine Learning	Ch 7, appendix E
The Elements of Statistical Learning	Sec. 4.5, Ch 12
Tutorial on Lagrange Multipliers	link