Center for Data Science
NYU

DS-GA 3001.007
Introduction to Machine Learning

Lecture 11

Support Vector Machines and Kernels

Classifying with Margins through Hinge Loss function

DS-GA 3001.007 Introduction to Machine Learning

Replacing features with relationships between features

Lecture 11

Support Vector Machines and Kernels

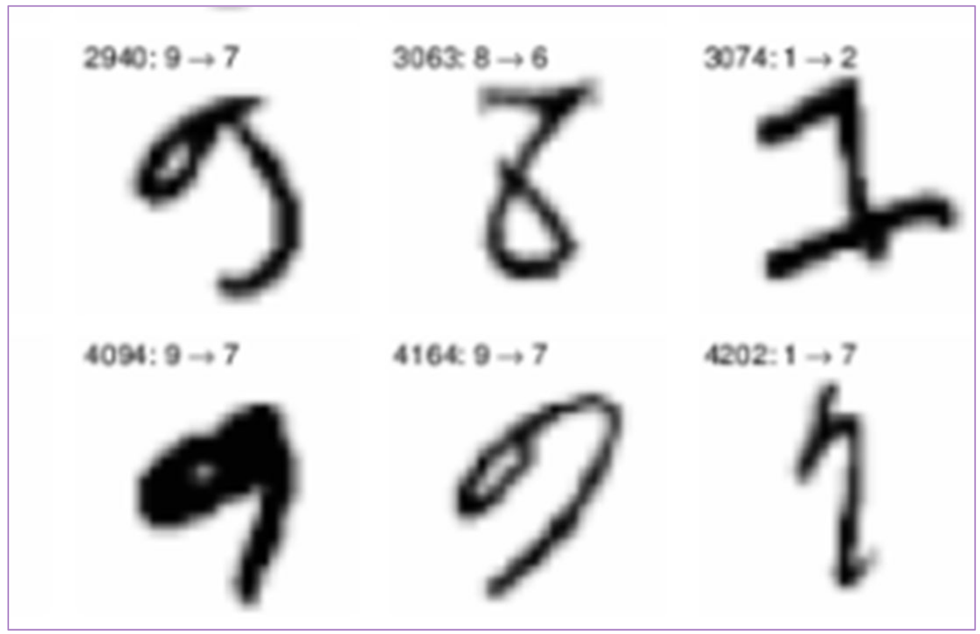
Announcements

- Homework
  - Homework 4 extended to **Wednesday November 13** at 11:59pm
  - Homework 5 due **Tuesday November 26** at 11:59pm
- Project
  - Milestone due **Thursday November 28** at 11:59pm
  - Background and Plans
- Labs
  - Submit on Jupyter Hub under Assignments tab

Refer to weekly agenda for more information

Homework 5

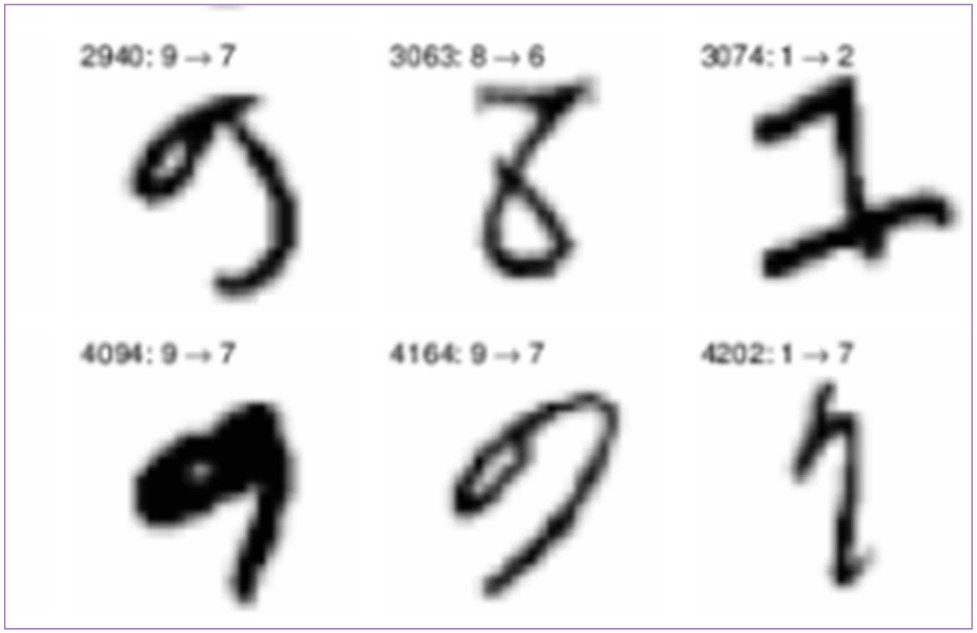
- Kernel Methods
  - New Kernels from Old Kernels
  - Modifying Pegasos Algorithm for Kernelized SVM
  - Image Classification
    - MNIST dataset of handwritten characters
    - 10 categories not 2 categories



Homework 5

- Kernel Methods
  - New Kernels from Old Kernels
  - Modifying Pegasos Algorithm for Kernelized SVM
  - Image Classification
    - MNIST dataset of handwritten characters
    - 10 categories not 2 categories

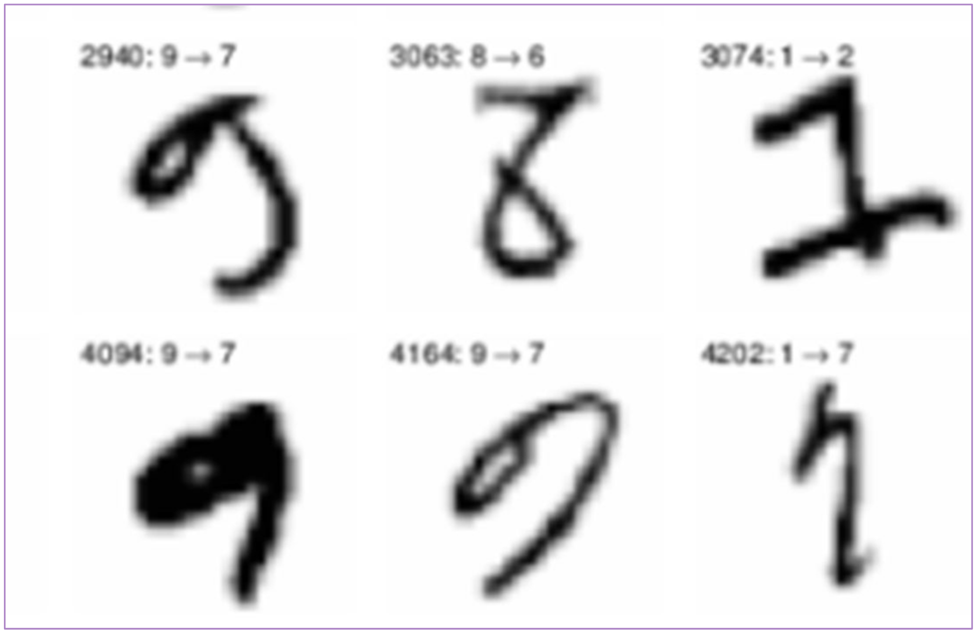
To check that kernel corresponds to relationship between features we would need to determine the feature transformation…



Homework 5

- Kernel Methods
  - New Kernels from Old Kernels
  - Modifying Pegasos Algorithm for Kernelized SVM
  - Image Classification
    - MNIST dataset of handwritten characters
    - 10 categories not 2 categories

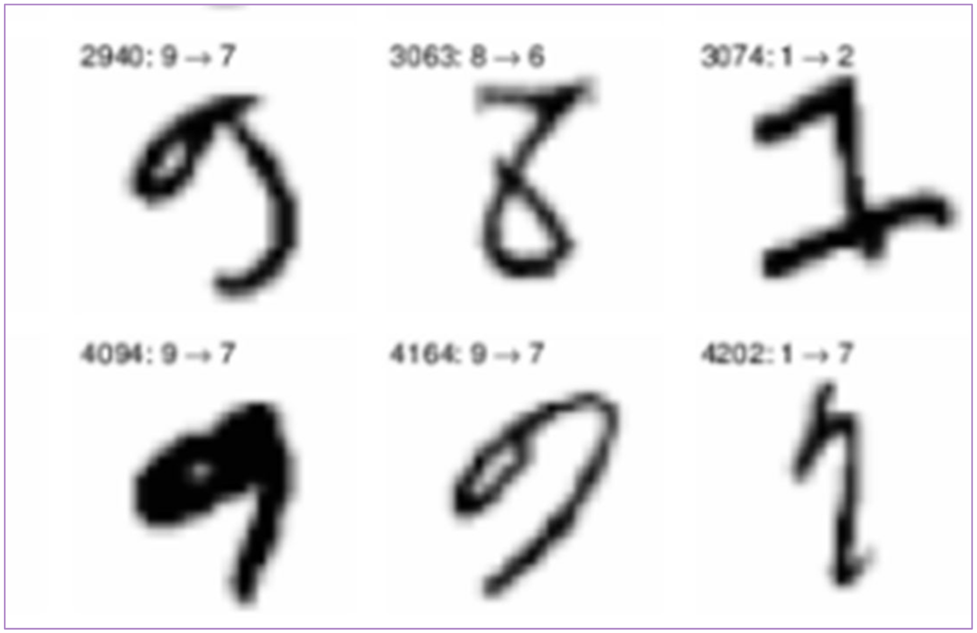
...instead express new kernel in terms of old kernel through scaling, sums and products



Homework 5

- Kernel Methods
  - New Kernels from Old Kernels
  - Modifying Pegasos Algorithm for Kernelized SVM
  - Image Classification
    - MNIST dataset of handwritten characters
    - 10 categories not 2 categories

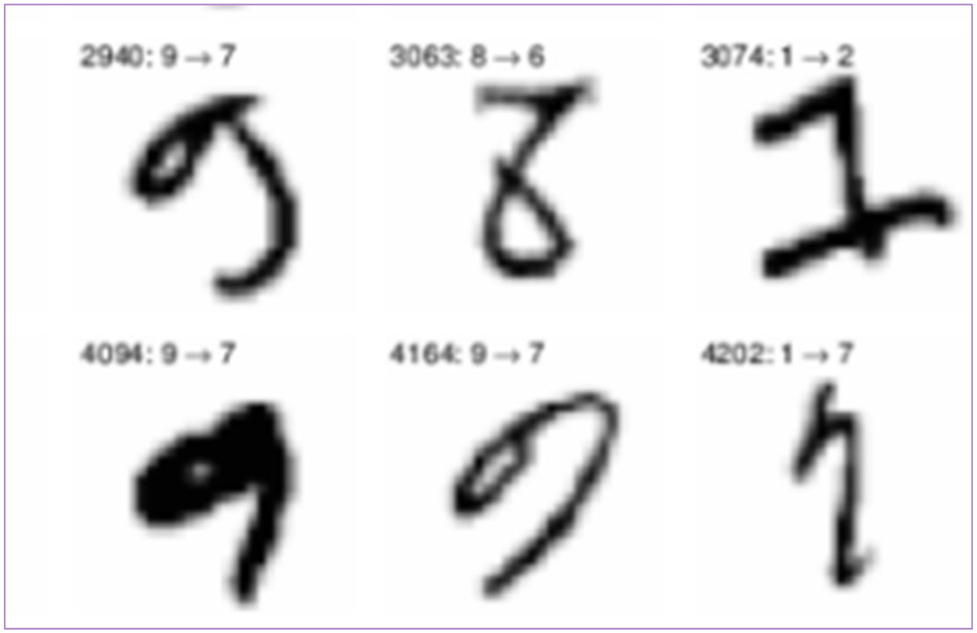
Pegasos is Stochastic Subgradient descent for Hinge Loss and L2 regularization with decreasing learning rate.



Homework 5

- Kernel Methods
  - New Kernels from Old Kernels
  - Modifying Pegasos Algorithm for Kernelized SVM
  - Image Classification
    - MNIST dataset of handwritten characters
    - 10 categories not 2 categories

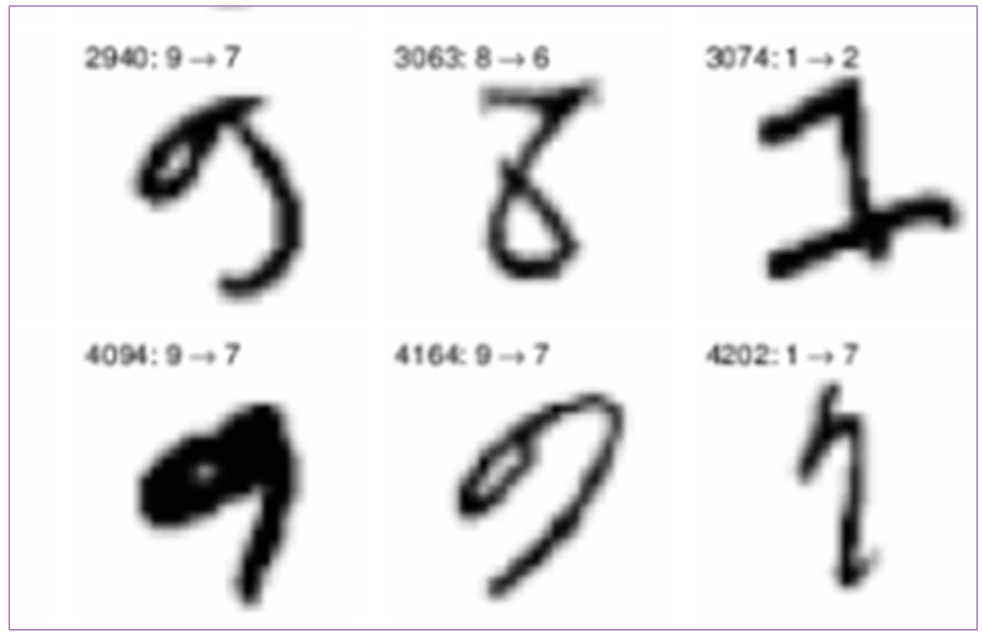
Kernelized SVM means expressing the weights in different way…



Homework 5

- Kernel Methods
  - New Kernels from Old Kernels
  - Modifying Pegasos Algorithm for Kernelized SVM
  - Image Classification
    - MNIST dataset of handwritten characters
    - 10 categories not 2 categories

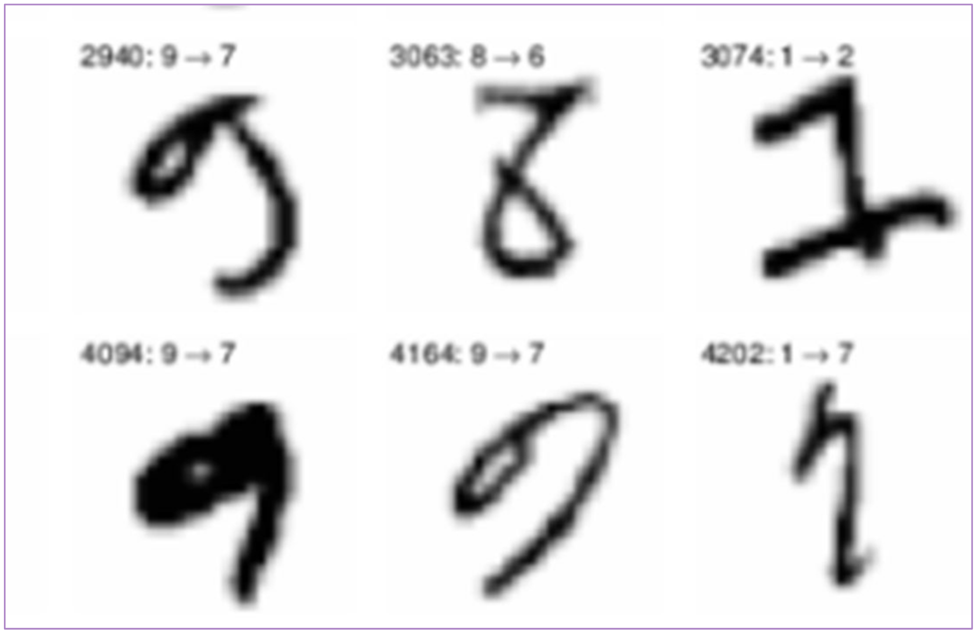
…you will determine the different update step in the algorithm to change your implementation



Homework 5

- Kernel Methods
  - New Kernels from Old Kernels
  - Modifying Pegasos Algorithm for Kernelized SVM
  - Image Classification
    - MNIST dataset of handwritten characters
    - 10 categories not 2 categories

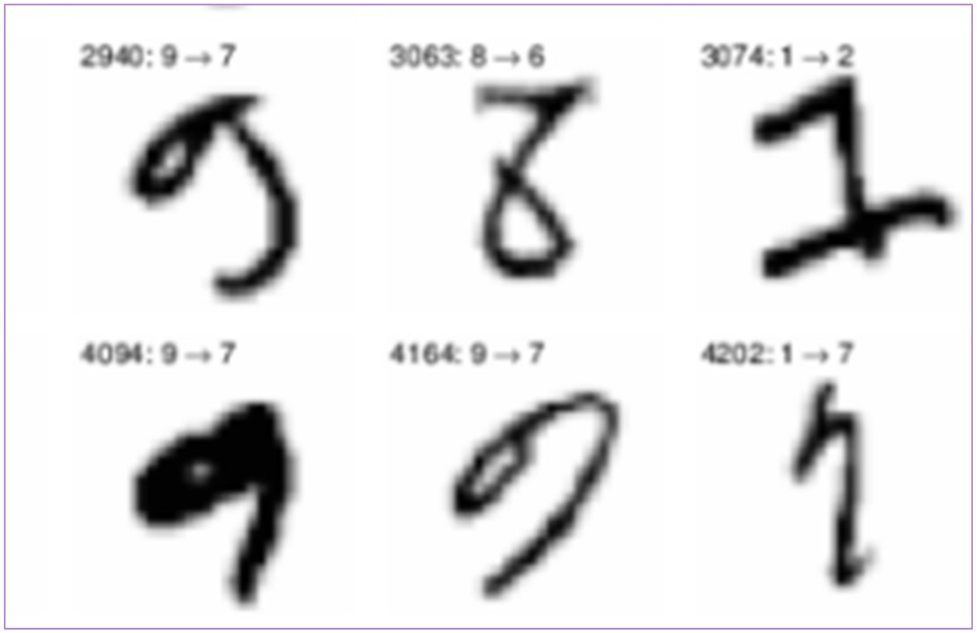
…you will determine the different update step in the algorithm to change your implementation



Homework 5

- Kernel Methods
  - New Kernels from Old Kernels
  - Modifying Pegasos Algorithm for Kernelized SVM
  - Image Classification
    - MNIST dataset of handwritten characters
    - 10 categories not 2 categories

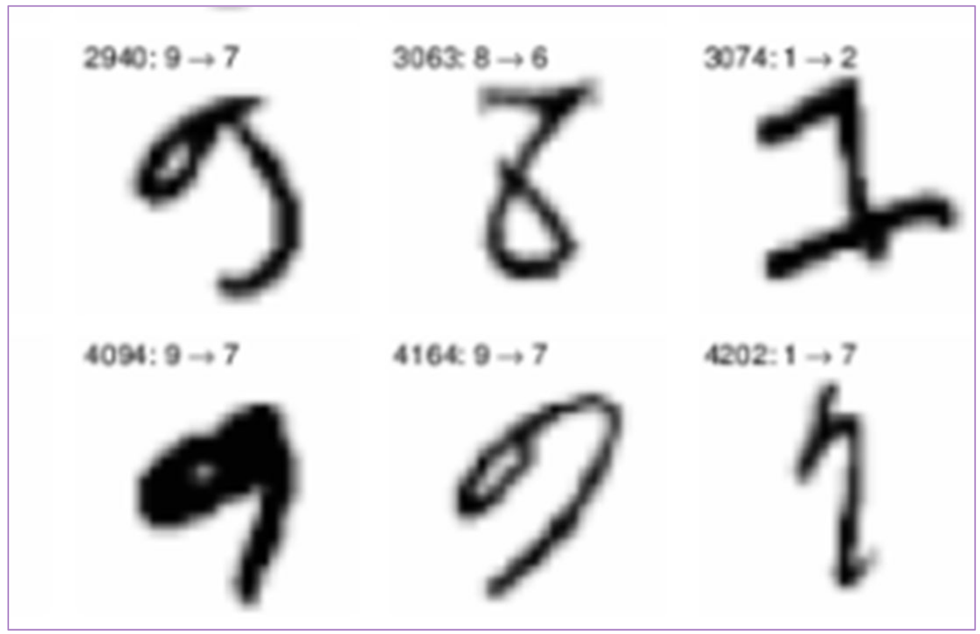
Training set 2000 handwritten digits. Test set 1000 handwritten digits. First column is label. Feature is 28x28 greyscale image.



Homework 5

- Kernel Methods
  - New Kernels from Old Kernels
  - Modifying Pegasos Algorithm for Kernelized SVM
  - Image Classification
    - MNIST dataset of handwritten characters
    - 10 categories not 2 categories

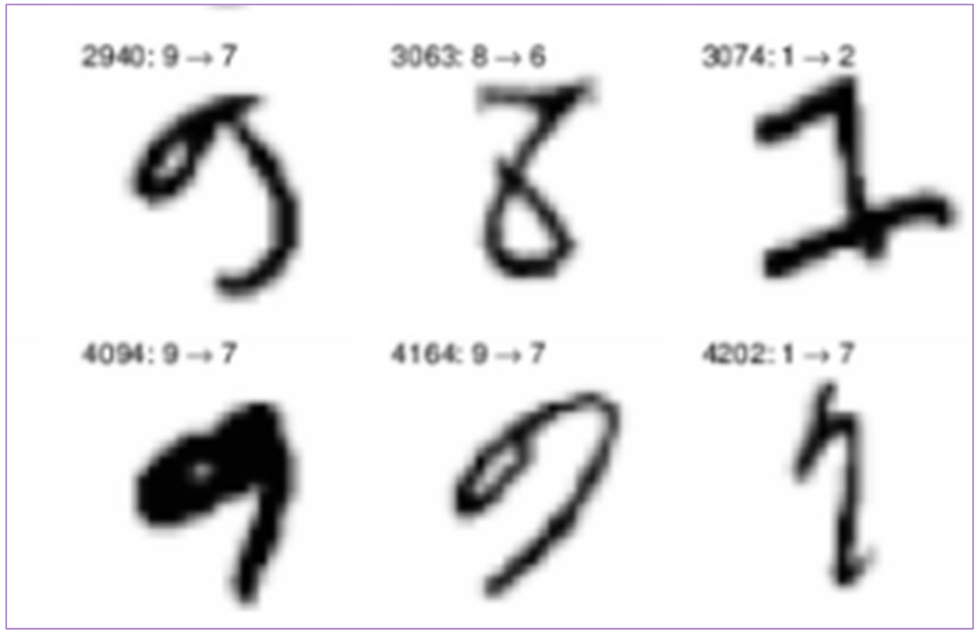
Classify into multiple categories…we will discuss one-vs-all classification next week



Homework 5

- Kernel Methods
  - New Kernels from Old Kernels
  - Modifying Pegasos Algorithm for Kernelized SVM
  - Image Classification
    - MNIST dataset of handwritten characters
    - 10 categories not 2 categories

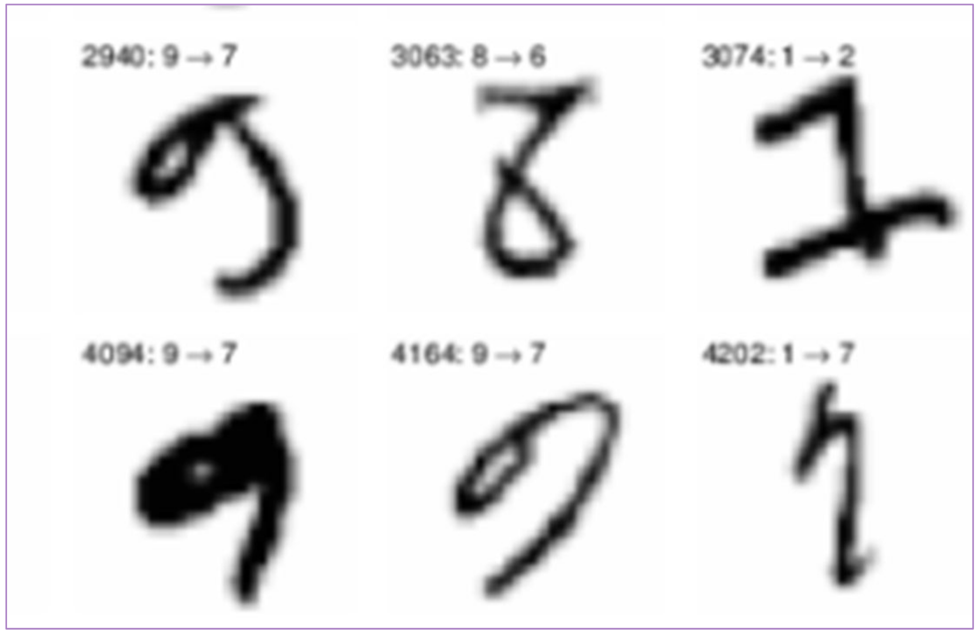
You will use your implementation of Pegasos and libSVM in sklearn that implements SMO…



…think of SMO as line search for quadratic programming problems because it changes variables two at a time.

Homework 5

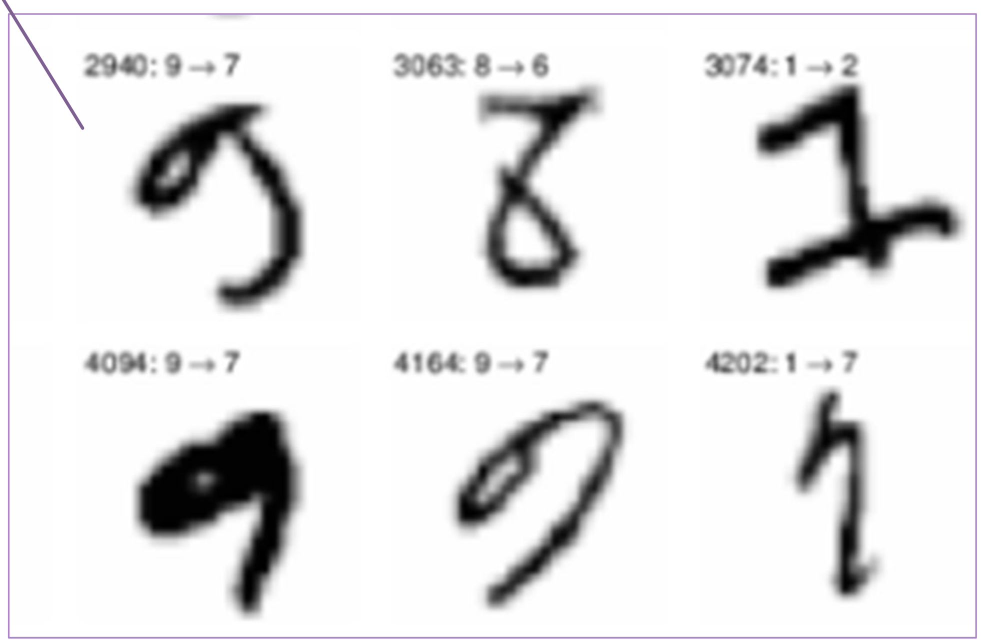
- Kernel Methods
  - New Kernels from Old Kernels
  - Modifying Pegasos Algorithm for Kernelized SVM
  - Image Classification
    - MNIST dataset of handwritten characters
    - 10 categories not 2 categories



Homework 5

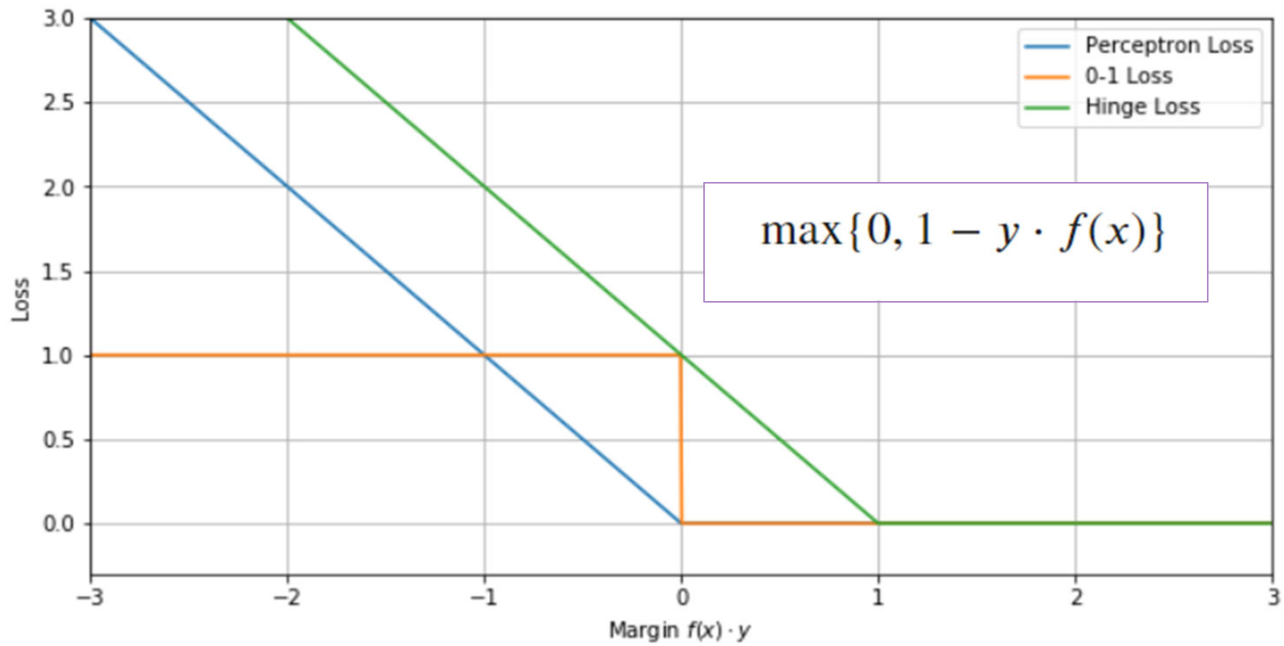
- Kernel Methods
  - New Kernels from Old Kernels
  - Modifying Pegasos Algorithm for Kernelized SVM
  - Image Classification
    - MNIST dataset of handwritten characters
    - 10 categories not 2 categories

http://scs.ryerson.ca/~aharley/vis/conv/flat.html



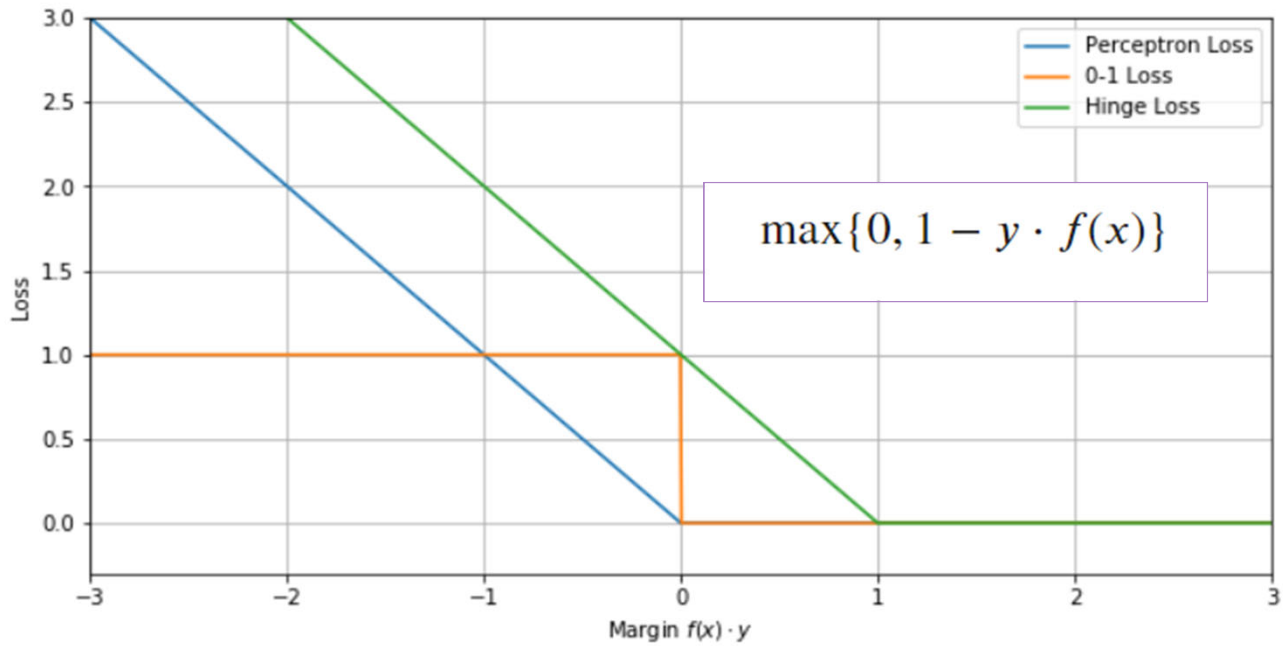
Review

- Minimizing Loss Functions
  - The empirical risk function might lead to optimization errors
  - Approaches
    - Find a replacement loss function



Review

- Minimizing Loss Functions
  - The empirical risk function might lead to optimization errors
  - Approaches
    - Find a replacement loss function
    - Determine another definition for gradient that handles corners

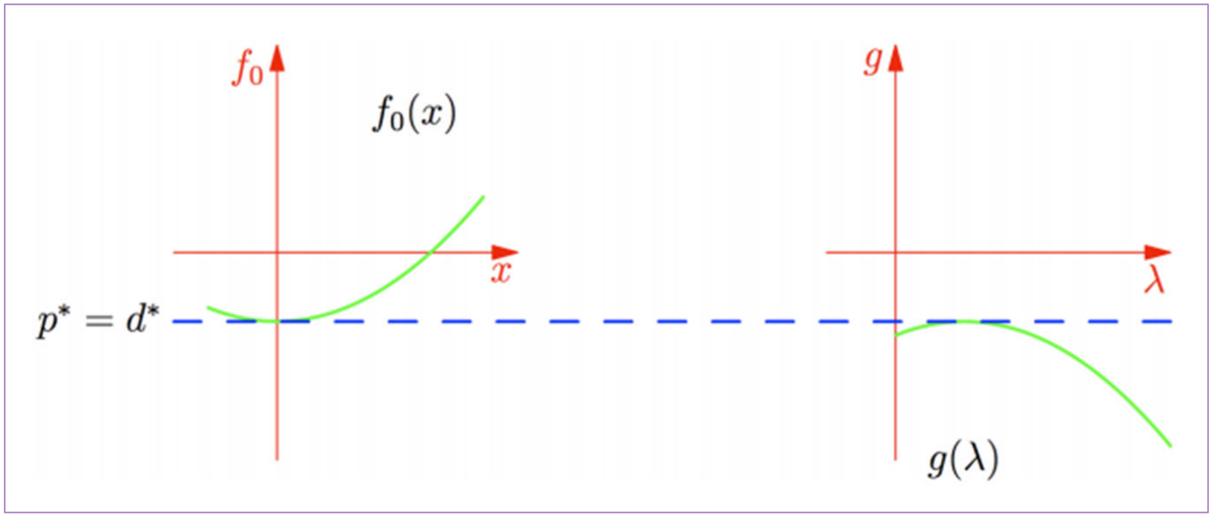


Derivative of hinge loss $\ell(m) = \max(0, 1 - m)$:

$$\ell'(m) = \begin{cases} 0 & m > 1 \\ -1 & m < 1 \\ \text{undefined} & m = 1 \end{cases}$$

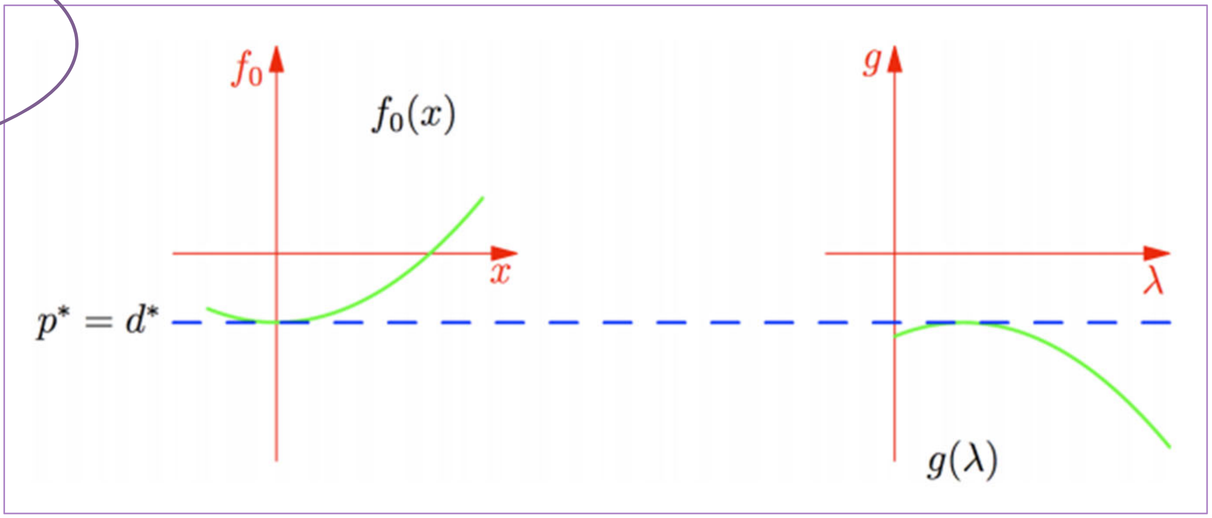
Review

- Minimizing Loss Functions
  - The empirical risk function might lead to optimization errors
  - Approaches
    - Rearrange optimization problem
      - Find equivalent problem
      - Combine objective and constraint
      - Switch order of minimization / maximization



Review

- Minimizing Loss Functions
  - The empirical risk function might lead to optimization errors
  - Approaches
    - Rearrange optimization problem
      - Find equivalent problem
      - Combine objective and constraint
      - Switch order of minimization / maximization



Rearranging Optimization Problems

Question

Suppose we have two functions $f : \mathbf{R}^d \to \mathbf{R}$ and $g : \mathbf{R}^d \to \mathbf{R}$. Now consider the following optimization problem:

$$\boxed{\min_{x \in \mathbf{R}^d} f(x) + g(x).}$$

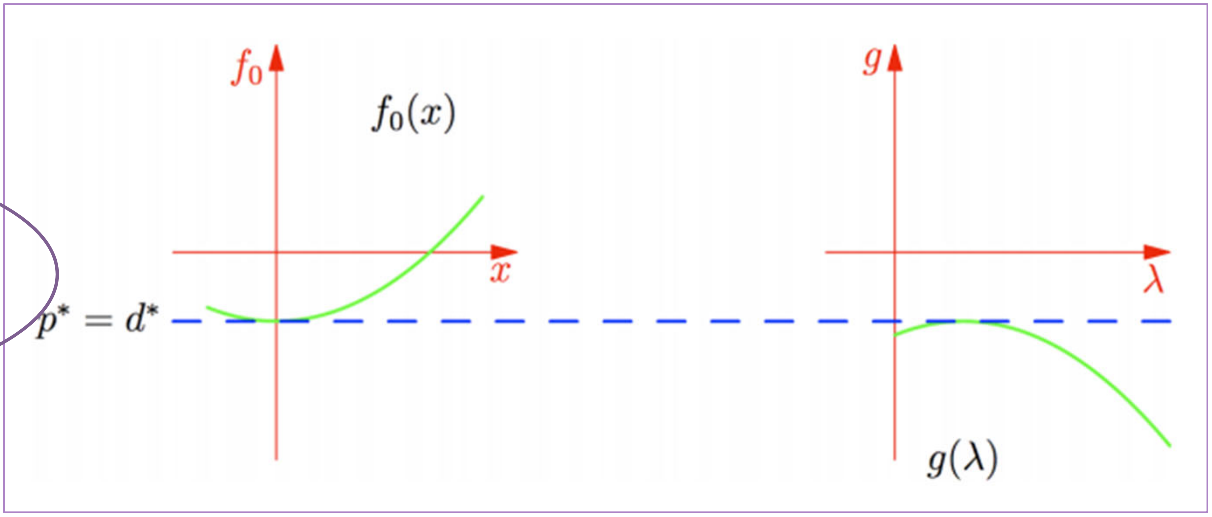
This is an unconstrained optimization problem. Let's also consider the following constrained optimization problem:

$$\boxed{\begin{array}{ll} \text{minimize} & f(x) + \xi \\ \text{subject to} & \xi \geq g(x). \end{array}}$$

Need to go both ways to have equivalent problem…there cannot be a gap

Review

- Minimizing Loss Functions
  - The empirical risk function might lead to optimization errors
  - Approaches
    - Rearrange optimization problem
      - Find equivalent problem
      - Combine objective and constraint
      - Switch order of minimization / maximization



Review

Question

Each row is a student strategy

$$A = \begin{bmatrix} 5 & 5 & 5 & 5 & 5 \\ 8 & 8 & 1 & 8 & 8 \\ +\infty & +\infty & +\infty & 0 & +\infty \end{bmatrix}$$

The entries represent points lost on the assignment

Each column is a grader strategy

Review

Question

$$A = \begin{bmatrix} 5 & 5 & 5 & 5 & 5 \\ 8 & 8 & 1 & 8 & 8 \\ +\infty & +\infty & +\infty & 0 & +\infty \end{bmatrix}$$

- We always have

$$\boxed{\max_j \min_i a_{ij} = d^* \leq p^* = \min_i \max_j a_{ij}.}$$

$$p^* = \min_i \max_j a_{ij}$$

$$d^* = \max_j \min_i a_{ij}$$

Review

Example

$$A = \begin{bmatrix} 5 & 5 & 5 & 5 & 5 \\ 8 & 8 & 1 & 8 & 8 \\ +\infty & +\infty & +\infty & 0 & +\infty \end{bmatrix}$$

- We always have

$$\boxed{\max_j \min_i a_{ij} = d^* \leq p^* = \min_i \max_j a_{ij}.}$$

$$p^* = \min_i \max_j a_{ij}$$

$$d^* = \max_j \min_i a_{ij}$$

because

$$d^* = a_{i_d j_d} \leq a_{i_p j_d} \leq a_{i_p j_p} = p^*.$$

Review

Example

$$A = \begin{bmatrix} 5 & 5 & 5 & 5 & 5 \\ 8 & 8 & 1 & 8 & 8 \\ +\infty & +\infty & +\infty & 0 & +\infty \end{bmatrix}$$

Primal Problem and Dual Problem may not be equal meaning you cannot switch max and min

$$p^* = \min_i \max_j a_{ij}$$

$$d^* = \max_j \min_i a_{ij}$$

- We always have

$$\boxed{\max_j \min_i a_{ij} = d^* \leq p^* = \min_i \max_j a_{ij}.}$$

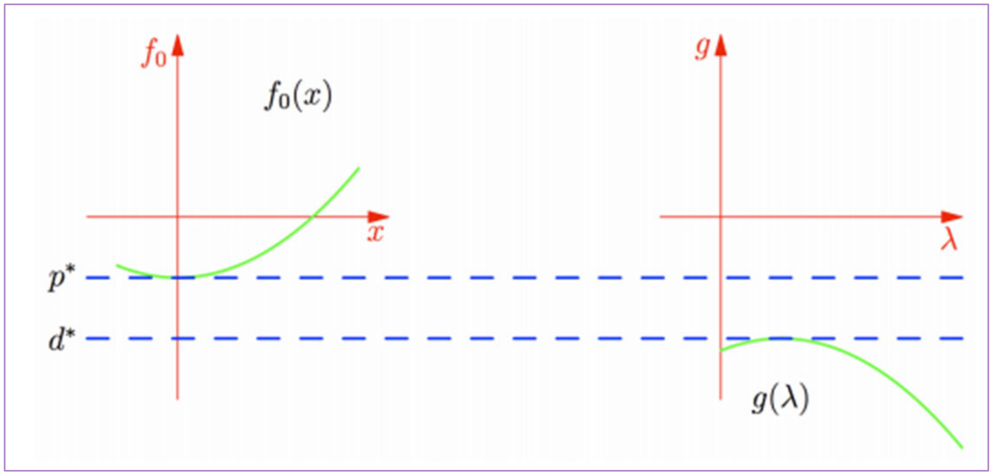
because

$$d^* = a_{i_d j_d} \leq a_{i_p j_d} \leq a_{i_p j_p} = p^*.$$

Review

The combination is the ***Lagrangian***

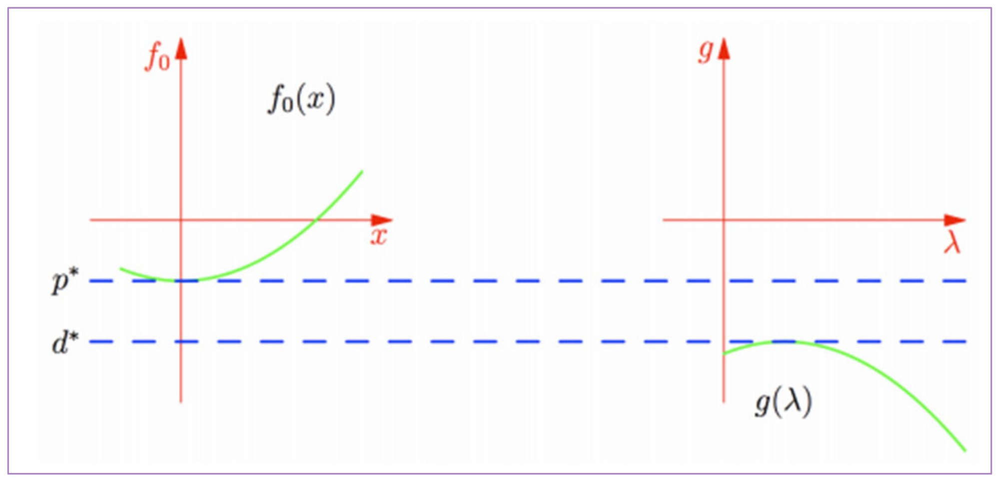
- Minimizing Loss Functions
  - The empirical risk function might lead to optimization errors
  - Approaches
    - Rearrange optimization problem
      - Find equivalent problem
      - Combine objective and constraint
      - Switch order of minimization / maximization



Review

Switching between primal and dual is called ***Lagrangian duality***

- Minimizing Loss Functions
  - The empirical risk function might lead to optimization errors
  - Approaches
    - Rearrange optimization problem
      - Find equivalent problem
      - Combine objective and constraint
      - Switch order of minimization / maximization



Review

Question

- Minimize $x + y$ subject to constraint $x^2 + y^2 = 1$

Review

Question

- Minimize $x + y$ subject to constraint $x^2 + y^2 = 1$
- We can combine the objective and constraint into a single function is Lagrangian

$$L(x, y, \lambda) = x + y + \lambda(x^2 + y^2 - 1)$$

We call this the ***Lagrange multiplier***…or ***dual variable*** in the context of Lagrangian duality

Review

Question

- Minimize $x + y$ subject to constraint $x^2 + y^2 = 1$
- We can combine the objective and constraint into a single function called the Lagrangian

$$\boxed{L(x, y, \lambda) = x + y + \lambda(x^2 + y^2 - 1)}$$

- Take derivative to find minimum

$$\nabla L = \begin{pmatrix} \frac{\partial L}{\partial x} \\ \frac{\partial L}{\partial y} \\ \frac{\partial L}{\partial \lambda} \end{pmatrix} = \begin{pmatrix} 1 + 2\lambda x \\ 1 + 2\lambda y \\ x^2 + y^2 - 1 \end{pmatrix}$$

Review

Question

- Minimize $x + y$ subject to constraint $x^2 + y^2 = 1$
- We can combine the objective and constraint into a single function called the Lagrangian

$$\boxed{L(x, y, \lambda) = x + y + \lambda(x^2 + y^2 - 1)}$$

- Take derivative to find minimum

$$\nabla L = \begin{pmatrix} \frac{\partial L}{\partial x} \\ \frac{\partial L}{\partial y} \\ \frac{\partial L}{\partial \lambda} \end{pmatrix} = \begin{pmatrix} 1 + 2\lambda x \\ 1 + 2\lambda y \\ x^2 + y^2 - 1 \end{pmatrix}$$

- Solutions at

$$\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) \text{ and } \left(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right).$$

Review

Question

- Minimize $x + y$ subject to constraint $x^2 + y^2 \leq 1$

Review

Question

- Minimize $x + y$ subject to constraint $x^2 + y^2 \leq 1$
- We can combine the objective and constraint into a single function the Lagrangian

$$\boxed{L(x, y, \lambda) = x + y + \lambda\left(1 - (x^2 + y^2)\right)}$$

- Here $\lambda > 0$

Review

Question

- Minimize $x + y$ subject to constraint $x^2 + y^2 \leq 1$
- We can combine the objective and constraint into a single function called the Lagrangian

$$L(x, y, \lambda) = x + y + \lambda\left(1 - (x^2 + y^2)\right)$$

- Here $\lambda > 0$

Take max over the dual variables and min over the primal variables

Question

Review

- Minimize $x + y$ subject to constraint $x^2 + y^2 \leq 1$
- We can combine the objective and constraint into a single function called the Lagrangian

$$\boxed{L(x, y, \lambda) = x + y + \lambda\left(1 - (x^2 + y^2)\right)}$$

- Here $\lambda > 0$

So penalization form and constraint form are definitely the same!

Review

Question

- Minimize $x + y$ subject to constraint $x^2 + y^2 \leq 1$
- We can combine the objective and constraint into a single function called the Lagrangian

$$\boxed{L(x, y, \lambda) = x + y + \lambda\left(1 - (x^2 + y^2)\right)}$$

- Here $\lambda > 0$. Take derivative to find minimum

$$\nabla L = \begin{pmatrix} \frac{\partial L}{\partial x} \\ \frac{\partial L}{\partial y} \end{pmatrix} = \begin{pmatrix} 1 + 2\lambda x \\ 1 + 2\lambda y \end{pmatrix}$$

Review

Question

- Minimize $x + y$ subject to constraint $x^2 + y^2 \leq 1$
- We can combine the objective and constraint into a single function called the Lagrangian

$$\boxed{L(x, y, \lambda) = x + y + \lambda\left(1 - (x^2 + y^2)\right)}$$

- Here $\lambda > 0$. Take derivative to find minimum

$$\nabla L = \begin{pmatrix} \frac{\partial L}{\partial x} \\ \frac{\partial L}{\partial y} \end{pmatrix} = \begin{pmatrix} 1 + 2\lambda x \\ 1 + 2\lambda y \end{pmatrix}$$

- Solutions at

$$\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) \text{ and } \left(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right).$$

Review

Question

- Minimize $x + y$ subject to constraint $x^2 + y^2 \leq 1$
- We can combine the objective and constraint into a single function called the Lagrangian

$$\boxed{L(x, y, \lambda) = x + y + \lambda\left(1 - (x^2 + y^2)\right)}$$

- Here $\lambda > 0$. Take derivative to find minimum

$$\nabla L = \begin{pmatrix} \frac{\partial L}{\partial x} \\ \frac{\partial L}{\partial y} \end{pmatrix} = \begin{pmatrix} 1 + 2\lambda x \\ 1 + 2\lambda y \end{pmatrix}$$

Note that solutions not unique. Is the objective convex? Is the objective concave?

- Solutions at

$$\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) \text{ and } \left(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right).$$

Review

Question

- Minimize $x + y$ subject to constraint $x^2 + y^2 \leq 1$
- We can combine the objective and constraint into a single function called the Lagrangian

$$\boxed{L(x, y, \lambda) = x + y + \lambda\left(1 - (x^2 + y^2)\right)}$$

- Here $\lambda > 0$. Take derivative to find minimum

$$\nabla L = \begin{pmatrix} \frac{\partial L}{\partial x} \\ \frac{\partial L}{\partial y} \end{pmatrix} = \begin{pmatrix} 1 + 2\lambda x \\ 1 + 2\lambda y \end{pmatrix}$$

- Solutions at

$$\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) \text{ and } \left(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right).$$

At the minimizer the constraint is satisfied…this is example of complementary slackness

Agenda

- Lesson
  - Support Vector Machines
    - Hard Margin
    - Dual problem
  - Kernels
    - Relationships between features
- Demo
  - libSVM package for SMO

Objectives

- What is the geometric interpretation of SVM?
- What insights can we gain from the dual formulation of SVM?
- Why would kernels be helpful with many features?
- **Readings:**
  - Shalev-Schwarz Chapter 16
  - Murphy Chapter 14.5 (see 14.5.2.4 for multiclass classification)

Agenda

- Lesson
  - Support Vector Machines
    - Hard Margin
    - Dual problem
  - Kernels
    - Relationships between features
- Demo
  - libSVM package for SMO

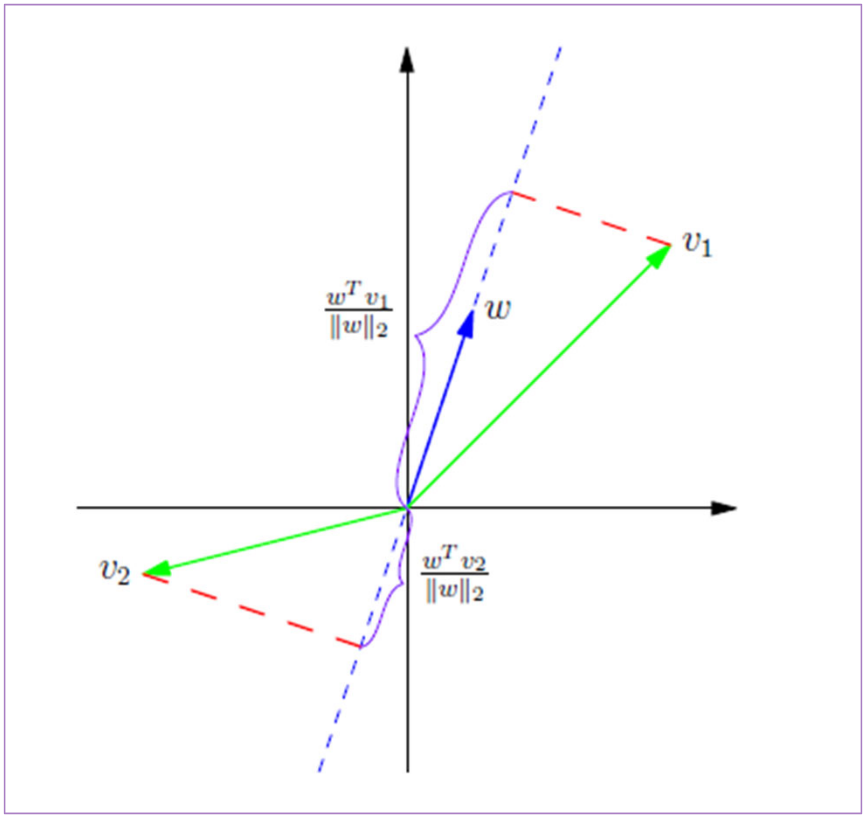
Objectives

- **What is the geometric interpretation of SVM?**
- What insights can we gain from the dual formulation of SVM?
- Why would kernels be helpful with many features?
- **Readings:**
  - Shalev-Schwarz Chapter 16
  - Murphy Chapter 14.5 (see 14.5.2.4 for multiclass classification)

Support Vector Machines

Hard Margin

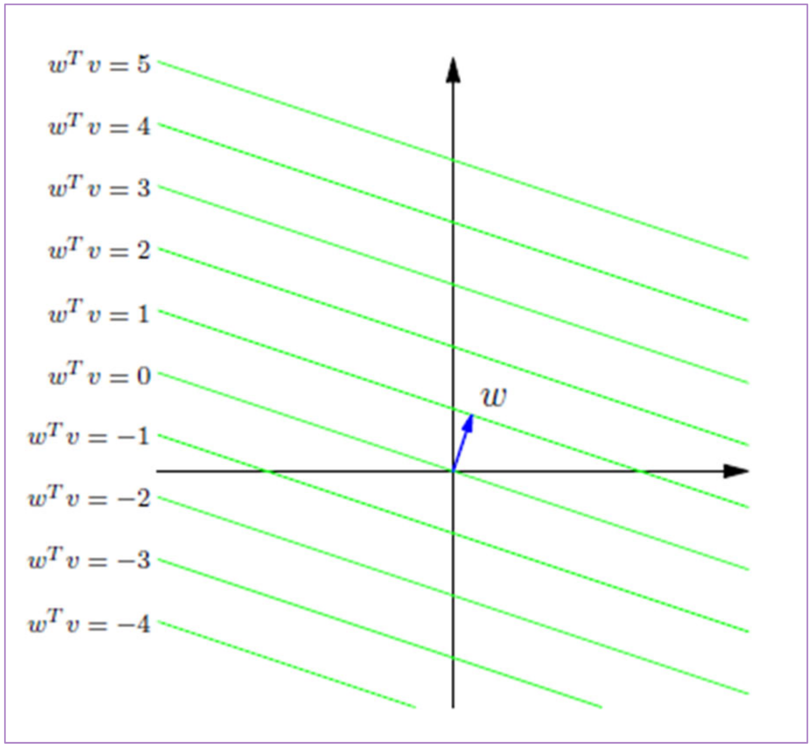
- A vector determines a plane in space. It consists of vectors at a right angle.
- Note that scaling the vector yields the same plane of perpendicular vectors



Support Vector Machines

Hard Margin

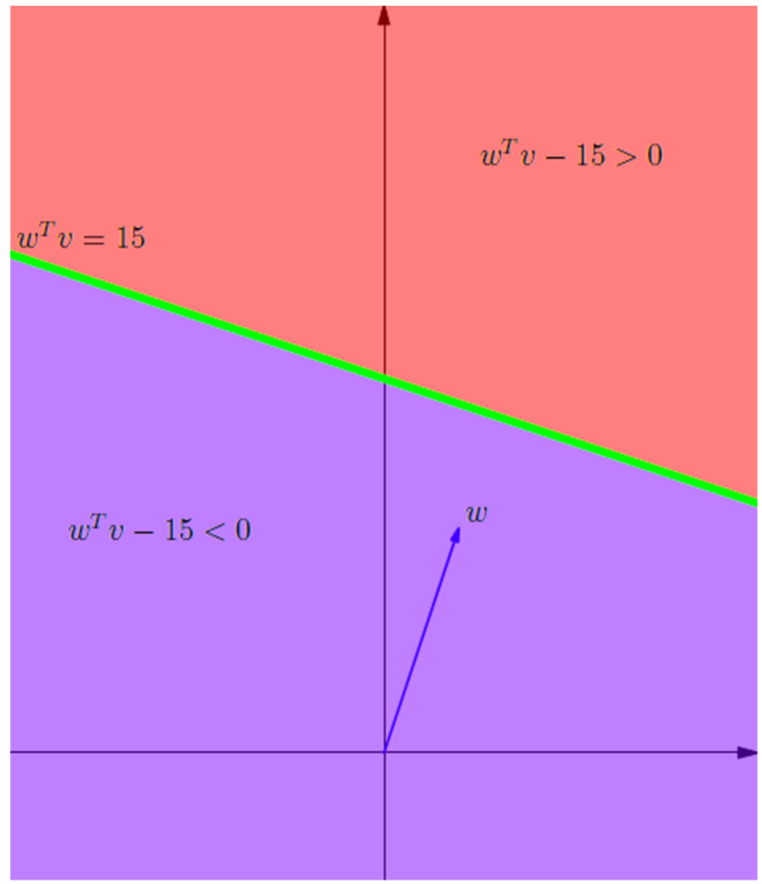
- A vector determines a plane in space. It consists of vectors at a right angle.
- Note that scaling the vector yields the same plane of perpendicular vectors
- The offset term b determine shift up or down



Hard Margin

Support Vector Machines

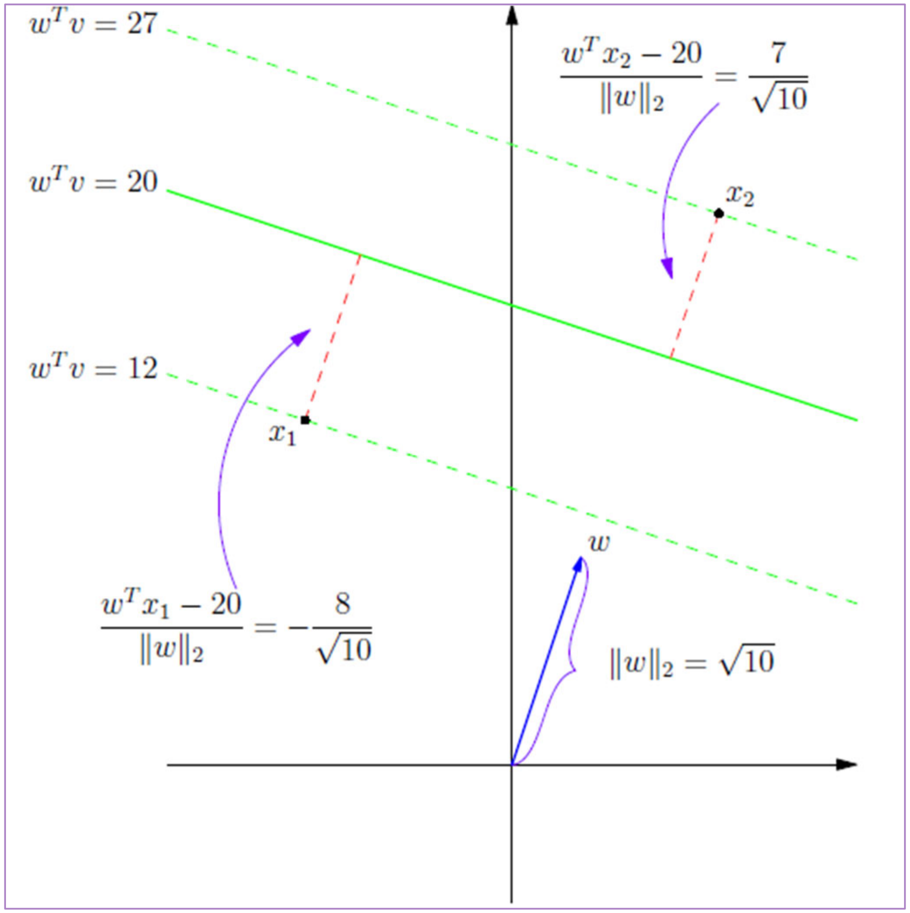
- Note that distance from the plane has a sign…
  - Positive for above the plane
  - Negative for below the plane



Support Vector Machines

Hard Margin

- Note that distance from the plane has a sign…
  - Positive for above the plane
  - Negative for below the plane
- The ***geometric margin*** gives the signed distance from the plane
  - Note the term in the denominator is different from the (functional) margin

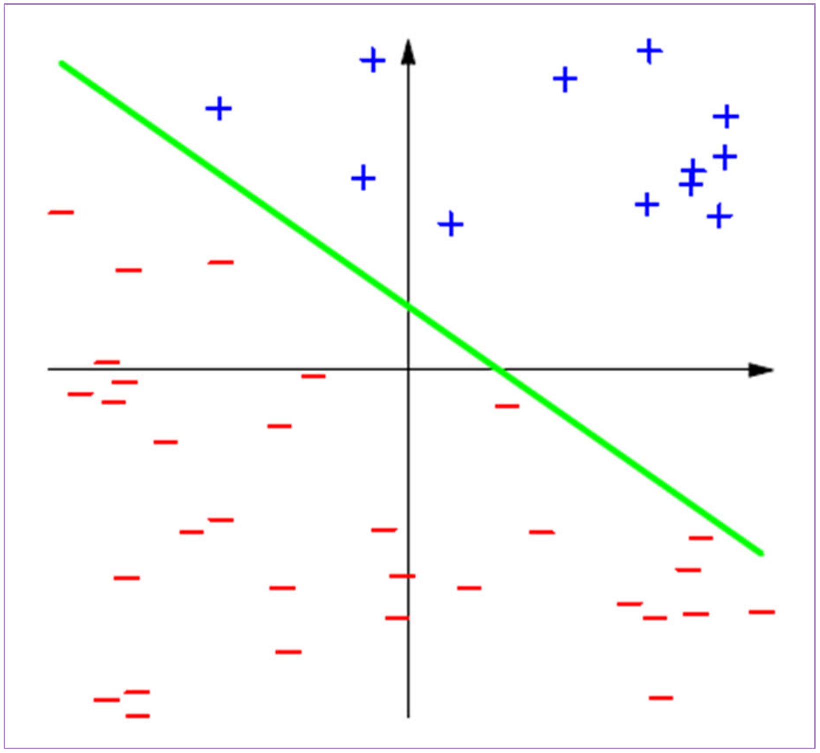


Support Vector Machines

Hard Margin

- We want to classify points in the training set with a separating plane.
  - Here the hypothesis set consists of functions

$$\mathrm{sgn}(w^T x + a).$$

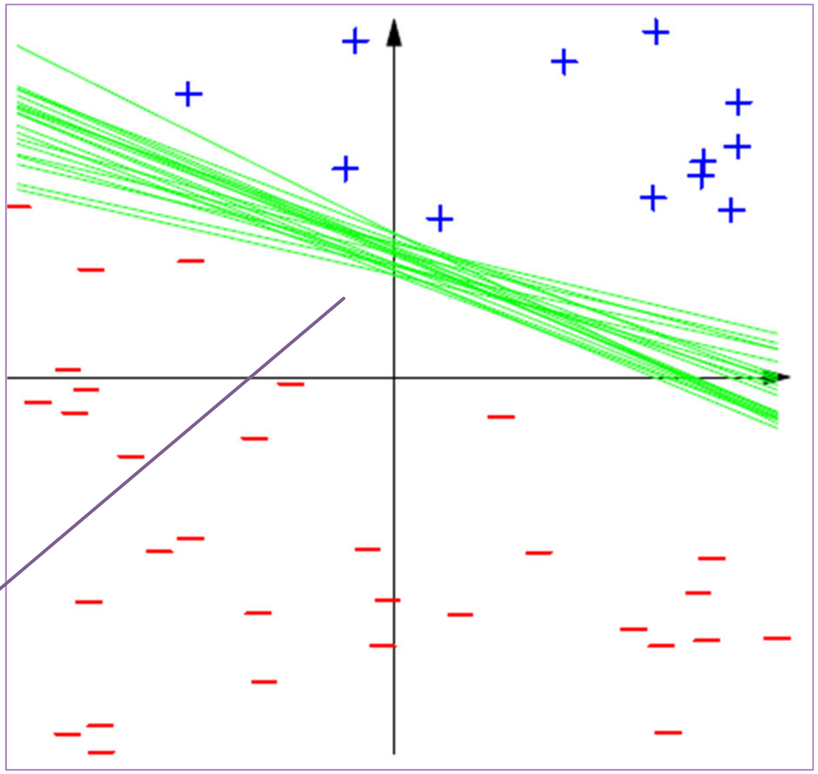


Support Vector Machines

Hard Margin

- We want to classify points in the training set with a separating plane.
  - Here the hypothesis set consists of functions

$$\mathrm{sgn}(w^T x + a).$$

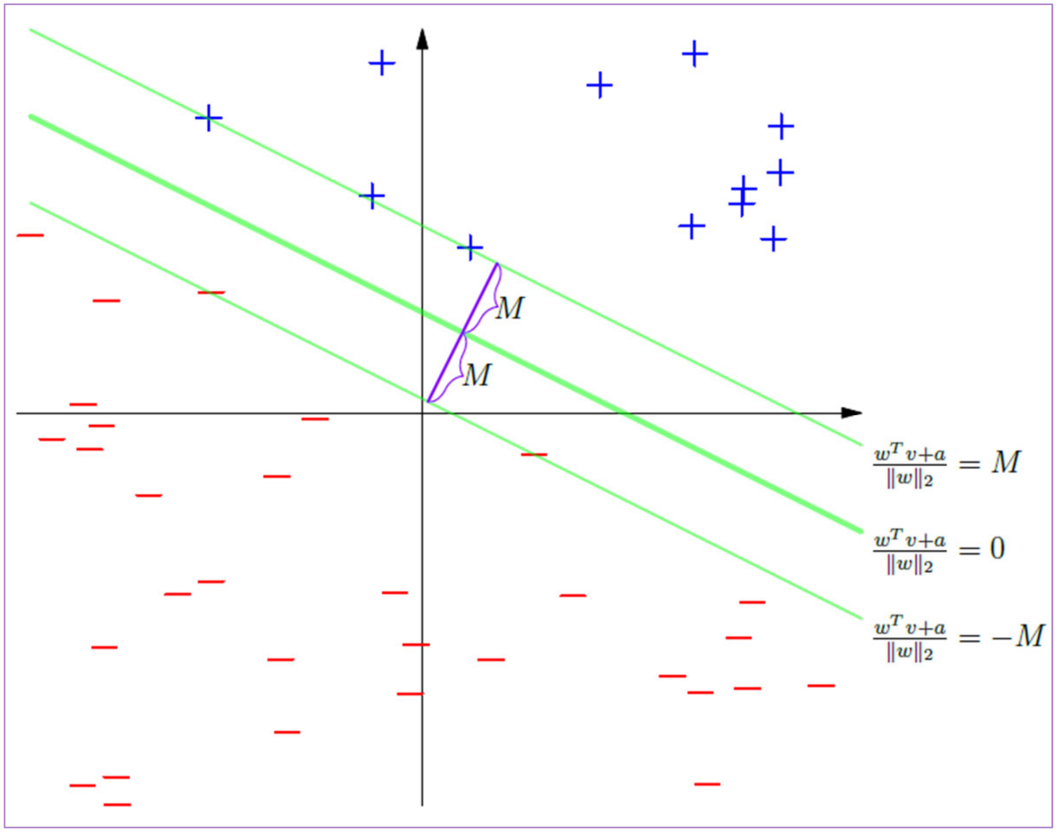


This was a problem with Perceptron…could lead to overfitting

Hard Margin

Support Vector Machines

- We want to classify points in the training set with a separating plane.
  - We can incorporate confidence into the hypotheses with the geometric margin

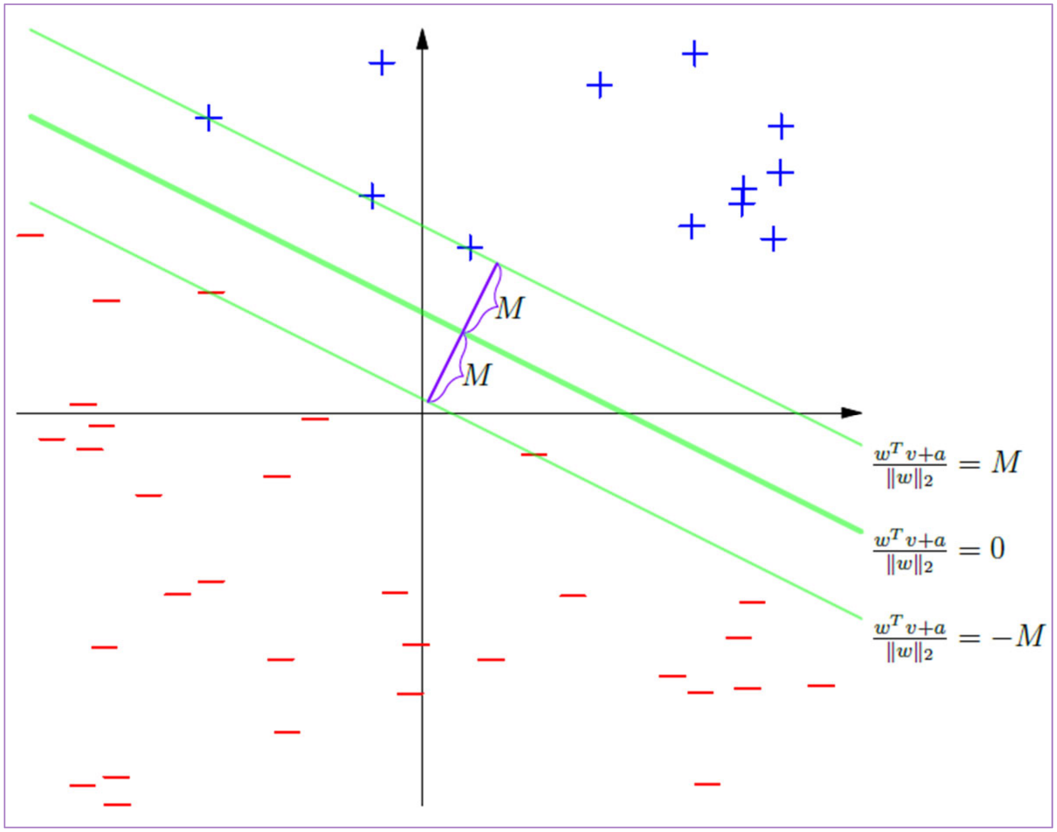


Support Vector Machines

Hard Margin

- We want to classify points in the training set with a separating plane.
  - We can incorporate confidence into the hypotheses with the geometric margin
  - Distance from plane is

$$\left| \frac{w^T x_i + a}{\|w\|_2} \right| = \frac{y_i(w^T x_i + a)}{\|w\|_2}$$



Support Vector Machines

Hard Margin

- We want to classify points in the training set with a separating plane.
  - We can incorporate confidence into the hypotheses with the geometric margin
  - Distance from plane is

$$\boxed{\left|\frac{w^T x_i + a}{\|w\|_2}\right| = \frac{y_i(w^T x_i + a)}{\|w\|_2}}$$

Goal

$$\text{maximize}_{w,a} \quad \min_i \frac{y_i(w^T x_i + a)}{\|w\|_2}.$$

Support Vector Machines

Hard Margin

- We want to classify points in the training set with a separating plane.
  - We can incorporate confidence into the hypotheses with the geometric margin
  - Distance from plane is

$$\boxed{\left|\frac{w^T x_i + a}{\|w\|_2}\right| = \frac{y_i(w^T x_i + a)}{\|w\|_2}}$$

Goal

$$\text{maximize}_{w,a} \quad \min_i \frac{y_i(w^T x_i + a)}{\|w\|_2}.$$

Rearrange to equivalent problem

$$\begin{aligned} &\text{maximize}_{w,a,M} && M \\ &\text{subject to} && \frac{y_i(w^T x_i + a)}{\|w\|_2} \geq M \quad \text{for all } i \end{aligned}$$

Support Vector Machines

Hard Margin

- We want to classify points in the training set with a separating plane.
  - We can incorporate confidence into the hypotheses with the geometric margin
  - Distance from plane is

$$\boxed{\left|\frac{w^T x_i + a}{\|w\|_2}\right| = \frac{y_i(w^T x_i + a)}{\|w\|_2}}$$

Goal

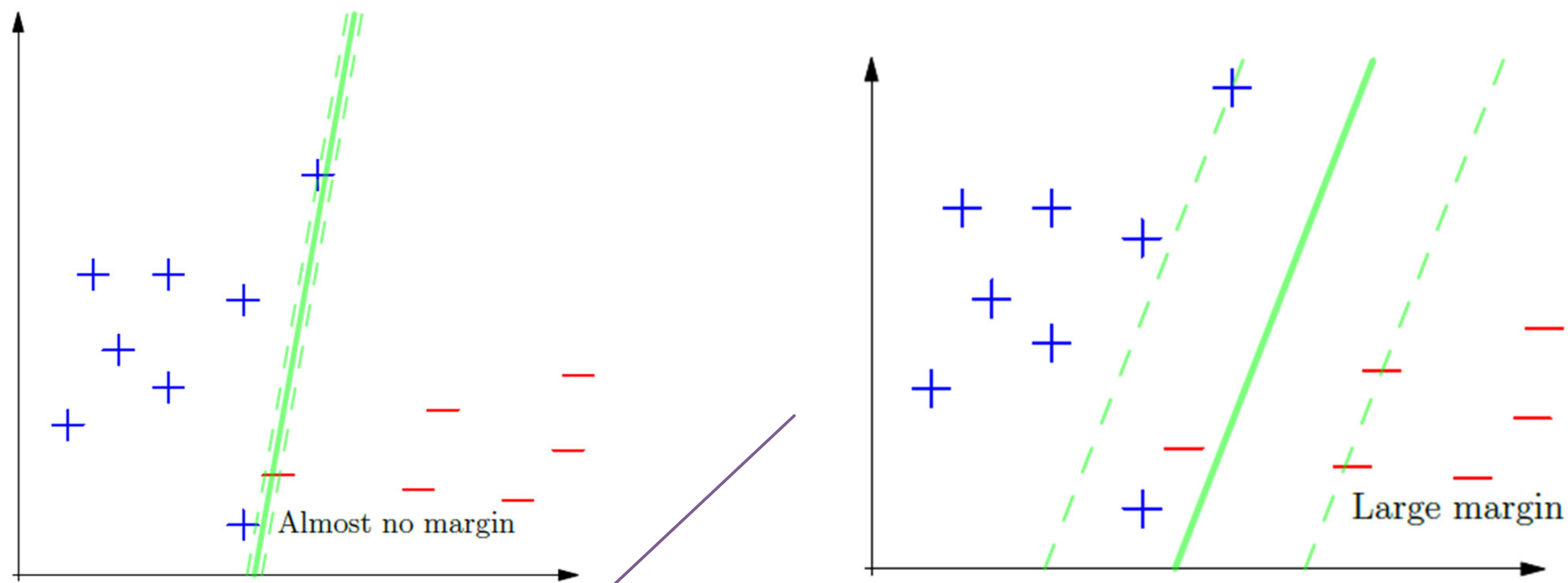
$$\begin{aligned} &\text{maximize}_{w,a,M} && M \\ &\text{subject to} && \frac{y_i(w^T x_i + a)}{\|w\|_2} \geq M \quad \text{for all } i \end{aligned}$$

Rearrange to equivalent problem

$$\begin{aligned} &\text{maximize}_{w,a} && 1/\|w\|_2 \\ &\text{subject to} && y_i(w^T x_i + a) \geq 1 \quad \text{for all } i \end{aligned}$$

Support Vector Machines

Hard Margin



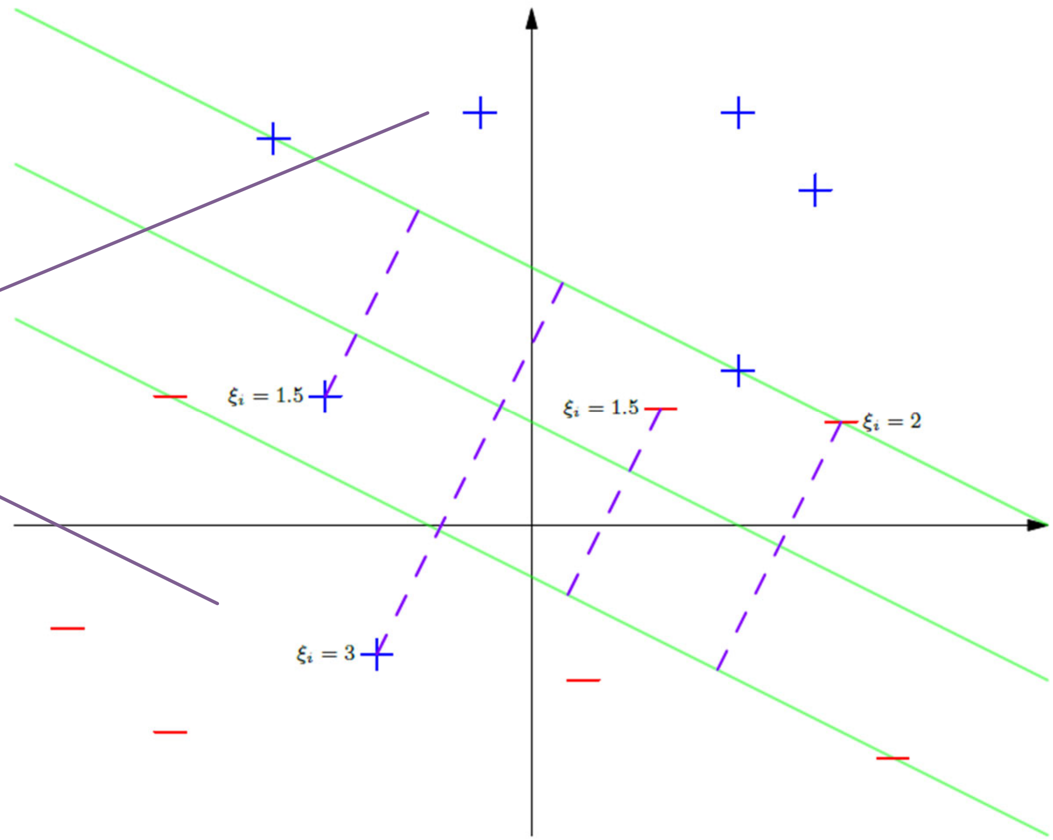
Which separating plan will better generalize from in-sample to out-of-sample?

Support Vector Machines

Soft Margin

Allow for violations for the constraint...instead use penalization

Slack variable is 0 for correctly classified points



Support Vector Machines

Soft Margin

- We can avoid overfitting through relaxation of the constraint.
- We switch from constraint form to penalization form

Goal

$$\begin{array}{ll} \text{minimize}_{w,a} & \|w\|_2^2 \\ \text{subject to} & y_i(w^T x_i + a) \geq 1 \quad \text{for all } i \end{array}$$

Rearrange to equivalent problem

$$\begin{array}{ll} \text{minimize}_{w,a,\xi} & \|w\|_2^2 + \frac{C}{n}\sum_{i=1}^n \xi_i \\ \text{subject to} & y_i(w^T x_i + a) \geq 1 - \xi_i \quad \text{for all } i \\ & \xi_i \geq 0 \quad \text{for all } i. \end{array}$$

Support Vector Machines

Soft Margin

$$\begin{aligned}
&\text{minimize} && \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i \\
&\text{subject to} && \xi_i \geqslant \left(1 - y_i \left[w^T x_i + b\right]\right) \text{ for } i = 1, \ldots, n \\
& && \xi_i \geqslant 0 \text{ for } i = 1, \ldots, n
\end{aligned}$$

Differentiable with n + d + 1 unknowns

Support Vector Machines

Soft Margin

$$\text{minimize} \quad \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i$$

$$\text{subject to} \quad \xi_i \geqslant \max\left(0, 1 - y_i\left[w^T x_i + b\right]\right).$$

$$\text{minimize} \quad \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i$$

$$\text{subject to} \quad \xi_i \geqslant \left(1 - y_i\left[w^T x_i + b\right]\right) \text{ for } i = 1, \ldots, n$$

$$\xi_i \geqslant 0 \text{ for } i = 1, \ldots, n$$

Differentiable with n + d + 1 unknowns

Support Vector Machines

Soft Margin

$$\min_{w \in \mathbf{R}^d, b \in \mathbf{R}} \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i\left[w^T x_i + b\right]\right).$$

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i \\ \text{subject to} & \xi_i \geqslant \max\left(0, 1 - y_i\left[w^T x_i + b\right]\right). \end{array}$$

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i \\ \text{subject to} & \xi_i \geqslant \left(1 - y_i\left[w^T x_i + b\right]\right) \text{ for } i = 1, \ldots, n \\ & \xi_i \geqslant 0 \text{ for } i = 1, \ldots, n \end{array}$$

Differentiable with n + d + 1 unknowns

Support Vector Machines

Soft Margin

$$\min_{w \in \mathbf{R}^d, b \in \mathbf{R}} \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i\left[w^T x_i + b\right]\right).$$

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i \\ \text{subject to} & \xi_i \geqslant \max\left(0, 1 - y_i\left[w^T x_i + b\right]\right). \end{array}$$

Quadratic Programming Problem…could solve with CVXOPT

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i \\ \text{subject to} & \xi_i \geqslant \left(1 - y_i\left[w^T x_i + b\right]\right) \text{ for } i = 1, \ldots, n \\ & \xi_i \geqslant 0 \text{ for } i = 1, \ldots, n \end{array}$$

Differentiable with n + d + 1 unknowns

Support Vector Machines

Soft Margin

$$\min_{w \in \mathbf{R}^d, b \in \mathbf{R}} \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i \left[w^T x_i + b\right]\right).$$

Support Vector Machines

Soft Margin

$$\min_{w \in \mathbf{R}^d, b \in \mathbf{R}} \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i\left[w^T x_i + b\right]\right).$$

Penalization form not constraint form with l2 regularization not l1 regularization

Support Vector Machines

Soft Margin

$$\min_{w \in \mathbf{R}^d, b \in \mathbf{R}} \frac{1}{2}||w||^2 + \frac{c}{n} \sum_{i=1}^{n} \max\left(0, 1 - y_i \left[w^T x_i + b\right]\right).$$

c not lambda

Penalization form not constraint form with l2 regularization not l1 regularization

Support Vector Machines

Soft Margin

b is intercept term in line…for classification with lines b is threshold

$$\min_{w \in \mathbf{R}^d, b \in \mathbf{R}} \frac{1}{2}||w||^2 + \frac{c}{n} \sum_{i=1}^{n} \max\left(0, 1 - y_i \left[w^T x_i + b\right]\right).$$

c not lambda

Penalization form not constraint form with l2 regularization not l1 regularization

Support Vector Machines

Soft Margin

b is intercept term in line…for classification with lines b is threshold

$$\min_{w \in \mathbf{R}^d, b \in \mathbf{R}} \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i\left[w^T x_i + b\right]\right).$$

c not lambda

Measures the confidence of the prediction, that is, the margin

Penalization form not constraint form with l2 regularization not l1 regularization

Measures the accuracy of the classification

Demo

- Support Vector Machine
  - Iris Dataset
  - Features
    - Petal Width
    - Petal Length
  - Classification
    - Iris-Versicolor
    - Iris-Setosa

Take-Aways

- Why is SVM affected by scaling?
- How can soft margin SVM be used to detect outliers?
- How does changing C affect the classification? What prevents against overfitting.
- How do we use SVM in sklearn?

Agenda

- Lesson
  - Support Vector Machines
    - Hard Margin
    - Dual problem
  - Kernels
    - Relationships between features
- Demo
  - libSVM package for SMO

Objectives

- What is the geometric interpretation of SVM?
- **What insights can we gain from the dual formulation of SVM?**
- Why would kernels be helpful with many features?
- **Readings:**
  - Shalev-Schwarz Chapter 16
  - Murphy Chapter 14.5 (see 14.5.2.4 for multiclass classification)

Support Vector Machines

Dual Problem

- If we form the Lagrangian, then we can incorporate maximization into the minimization problem

$$
\begin{array}{ll}
\text{minimize} & \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i \\
\text{subject to} & -\xi_i \leqslant 0 \text{ for } i = 1, \ldots, n \\
& \left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \leqslant 0 \text{ for } i = 1, \ldots, n
\end{array}
$$

Support Vector Machines

Dual Problem

- If we form the Lagrangian, then we can incorporate maximization into the minimization problem

$$
\begin{aligned}
\text{minimize} \quad & \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i \\
\text{subject to} \quad & -\xi_i \leqslant 0 \text{ for } i = 1,\ldots,n \\
& \left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \leqslant 0 \text{ for } i = 1,\ldots,n
\end{aligned}
$$

Lagrange Multiplier	Constraint
λ_i	$-\xi_i \leqslant 0$
α_i	$\left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \leqslant 0$

Dual Problem

Support Vector Machines

- If we form the Lagrangian, then we can incorporate maximization into the minimization problem

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & -\xi_i \leqslant 0 \text{ for } i = 1, \dots, n \\ & \left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \leqslant 0 \text{ for } i = 1, \dots, n \end{aligned}$$

Lagrange Multiplier	Constraint
λ_i	$-\xi_i \leqslant 0$
α_i	$\left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \leqslant 0$

$$L(w, b, \xi, \alpha, \lambda) = \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i \left(1 - y_i\left[w^T x_i + b\right] - \xi_i\right) + \sum_{i=1}^{n} \lambda_i (-\xi_i)$$

Support Vector Machines

Dual Problem

- If we form the Lagrangian, then we can incorporate maximization into the minimization problem

$$\begin{aligned} p^* &= \inf_{w,\xi,b} \sup_{\alpha,\lambda\succeq 0} L(w,b,\xi,\alpha,\lambda) \\ &\geqslant \sup_{\alpha,\lambda\succeq 0} \inf_{w,b,\xi} L(w,b,\xi,\alpha,\lambda) = d^* \end{aligned}$$

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & -\xi_i \leqslant 0 \text{ for } i = 1,\dots,n \\ & \left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \leqslant 0 \text{ for } i = 1,\dots,n \end{aligned}$$

Lagrange Multiplier	Constraint
λ_i	$-\xi_i \leqslant 0$
α_i	$\left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \leqslant 0$

$$L(w,b,\xi,\alpha,\lambda) = \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - y_i\left[w^T x_i + b\right] - \xi_i\right) + \sum_{i=1}^{n}\lambda_i(-\xi_i)$$

Support Vector Machines

Dual Problem

- While primal problem and dual problem may not be equivalent, we are able to check ***constraint qualifications*** that guarantee equivalence

$$
\begin{aligned}
p^* &= \inf_{w,\xi,b} \sup_{\alpha,\lambda \succeq 0} L(w,b,\xi,\alpha,\lambda) \\
&\geqslant \sup_{\alpha,\lambda \succeq 0} \inf_{w,b,\xi} L(w,b,\xi,\alpha,\lambda) = d^*
\end{aligned}
$$

$$
\begin{aligned}
\text{minimize} \quad & \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i \\
\text{subject to} \quad & -\xi_i \leqslant 0 \text{ for } i = 1,\ldots,n \\
& \left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \leqslant 0 \text{ for } i = 1,\ldots,n
\end{aligned}
$$

Lagrange Multiplier	Constraint
λ_i	$-\xi_i \leqslant 0$
α_i	$\left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \leqslant 0$

$$
L(w,b,\xi,\alpha,\lambda) = \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - y_i\left[w^T x_i + b\right] - \xi_i\right) + \sum_{i=1}^{n}\lambda_i(-\xi_i)
$$

Support Vector Machines

Dual Problem

- While primal problem and dual problem may not be equivalent, we are able to check ***constraint qualifications*** that guarantee equivalence

$$p^* = \inf_{w,\xi,b} \sup_{\alpha,\lambda\succeq 0} L(w,b,\xi,\alpha,\lambda)$$
$$= \sup_{\alpha,\lambda\succeq 0} \inf_{w,b,\xi} L(w,b,\xi,\alpha,\lambda) = d^*$$

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} & -\xi_i \leqslant 0 \text{ for } i = 1,\dots,n \\ & \left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \leqslant 0 \text{ for } i = 1,\dots,n \end{array}$$

Lagrange Multiplier	Constraint
λ_i	$-\xi_i \leqslant 0$
α_i	$\left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \leqslant 0$

$$L(w,b,\xi,\alpha,\lambda) = \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - y_i\left[w^T x_i + b\right] - \xi_i\right) + \sum_{i=1}^{n}\lambda_i(-\xi_i)$$

Support Vector Machines

Dual Problem

- With the dual problem, we have to maximize...

$$g(\alpha,\lambda) = \inf_{w,b,\xi} L(w,b,\xi,\alpha,\lambda)$$

$$= \inf_{w,b,\xi}\left[\frac{1}{2}w^Tw + \sum_{i=1}^{n}\xi_i\left(\frac{c}{n}-\alpha_i-\lambda_i\right)+\sum_{i=1}^{n}\alpha_i\left(1-y_i\left[w^Tx_i+b\right]\right)\right]$$

Support Vector Machines

Dual Problem

- With the dual problem, we have to maximize…

$$g(\alpha,\lambda) = \inf_{w,b,\xi} L(w,b,\xi,\alpha,\lambda)$$

$$= \inf_{w,b,\xi} \left[\frac{1}{2} w^T w + \sum_{i=1}^{n} \xi_i \left(\frac{c}{n} - \alpha_i - \lambda_i \right) + \sum_{i=1}^{n} \alpha_i \left(1 - y_i \left[w^T x_i + b \right] \right) \right]$$

- What does the derivative tell us about the minimum?
  - Note that the weight, offset and slack are unconstrainted so we have no trouble taking derivatives

Support Vector Machines

Dual Problem

- With the dual problem, we have to maximize…

$$g(\alpha,\lambda) = \inf_{w,b,\xi} L(w,b,\xi,\alpha,\lambda)$$

$$= \inf_{w,b,\xi}\left[\frac{1}{2}w^T w + \sum_{i=1}^{n}\xi_i\left(\frac{c}{n}-\alpha_i-\lambda_i\right)+\sum_{i=1}^{n}\alpha_i\left(1-y_i\left[w^T x_i+b\right]\right)\right]$$

- What does the derivative tell us about the minimum?
  - Note that the weight, offset and slack are unconstrainted so we have no trouble taking derivatives

$$\partial_w L = 0 \iff w - \sum_{i=1}^{n}\alpha_i y_i x_i = 0 \iff \boxed{w = \sum_{i=1}^{n}\alpha_i y_i x_i}$$

Support Vector Machines

Dual Problem

- With the dual problem, we have to maximize…

$$g(\alpha,\lambda) = \inf_{w,b,\xi} L(w,b,\xi,\alpha,\lambda)$$

$$= \inf_{w,b,\xi}\left[\frac{1}{2}w^T w + \sum_{i=1}^{n}\xi_i\left(\frac{c}{n}-\alpha_i-\lambda_i\right)+\sum_{i=1}^{n}\alpha_i\left(1-y_i\left[w^T x_i+b\right]\right)\right]$$

- What does the derivative tell us about the minimum?
  - Note that the weight, offset and slack are unconstrainted so we have no trouble taking derivatives

$$\partial_b L = 0 \iff -\sum_{i=1}^{n}\alpha_i y_i = 0 \iff \boxed{\sum_{i=1}^{n}\alpha_i y_i = 0}$$

Support Vector Machines

Dual Problem

- With the dual problem, we have to maximize...

$$g(\alpha, \lambda) = \inf_{w,b,\xi} L(w, b, \xi, \alpha, \lambda)$$

$$= \inf_{w,b,\xi} \left[\frac{1}{2} w^T w + \sum_{i=1}^{n} \xi_i \left(\frac{c}{n} - \alpha_i - \lambda_i \right) + \sum_{i=1}^{n} \alpha_i \left(1 - y_i \left[w^T x_i + b \right] \right) \right]$$

- What does the derivative tell us about the minimum?
  - Note that the weight, offset and slack are unconstrainted so we have no trouble taking derivatives

$$\partial_{\xi_i} L = 0 \iff \frac{c}{n} - \alpha_i - \lambda_i = 0 \iff \boxed{\alpha_i + \lambda_i = \frac{c}{n}}$$

Support Vector Machines

Dual Problem

- Putting it together we obtain, these first order conditions show us that the dual problem is equivalent to

$$
\begin{aligned}
\sup_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_j^T x_i \\
\text{s.t.} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \\
& \alpha_i \in \left[0, \frac{c}{n}\right] \quad i = 1, \dots, n.
\end{aligned}
$$

Support Vector Machines

Dual Problem

- Putting it together we obtain, these first order conditions show us that the dual problem is equivalent to

$$\begin{aligned} \sup_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_j{}^T x_i \\ \text{s.t.} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \\ & \alpha_i \in \left[0, \frac{c}{n}\right] \quad i = 1, \ldots, n. \end{aligned}$$

The weights come from the dual variables

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i.$$

Support Vector Machines

Dual Problem

- Putting it together we obtain, these first order conditions show us that the dual problem is equivalent to

The hyperparameter limits the size…actually most will be 0

$$\sup_{\alpha} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_j^T x_i$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\alpha_i \in \left[0, \frac{c}{n}\right] \quad i = 1, \ldots, n.$$

The weights come from the dual variables

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i.$$

Support Vector Machines

Dual Problem

- Remember that we have ***complementary slackness*** conditions that relate values of dual variables and constraints

$$\alpha_i^* \left(1 - y_i f^*(x_i) - \xi_i^*\right) = 0$$

$$\lambda_i^* \xi_i^* = \left(\frac{c}{n} - \alpha_i^*\right) \xi_i^* = 0$$

Support Vector Machines

Dual Problem

- Remember that we have ***complementary slackness*** conditions that relate values of dual variables and constraints

$$\alpha_i^* = 0 \implies y_i f^*(x_i) \geqslant 1$$
$$\alpha_i^* \in \left(0, \frac{c}{n}\right) \implies y_i f^*(x_i) = 1$$
$$\alpha_i^* = \frac{c}{n} \implies y_i f^*(x_i) \leqslant 1$$

$$y_i f^*(x_i) < 1 \implies \alpha_i^* = \frac{c}{n}$$
$$y_i f^*(x_i) = 1 \implies \alpha_i^* \in \left[0, \frac{c}{n}\right]$$
$$y_i f^*(x_i) > 1 \implies \alpha_i^* = 0$$

$$\alpha_i^* \left(1 - y_i f^*(x_i) - \xi_i^*\right) = 0$$
$$\lambda_i^* \xi_i^* = \left(\frac{c}{n} - \alpha_i^*\right) \xi_i^* = 0$$

When are the constraints ***active?***

Agenda

- Lesson
  - Support Vector Machines
    - Hard Margin
    - Dual problem
  - Kernels
    - Relationships between features
- Demo
  - libSVM package for SMO

Objectives

- What is the geometric interpretation of SVM?
- What insights can we gain from the dual formulation of SVM?
- Why would kernels be helpful with many features?
- **Readings:**
  - Shalev-Schwarz Chapter 16
  - Murphy Chapter 14.5 (see 14.5.2.4 for multiclass classification)

Kernels

- The dual problem for SVM just depends on the inner products of the points in the sample
- How could we change to other relationships besides the ***linear kernel***?

$$
\begin{aligned}
\sup_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j {x_j}^T x_i \\
\text{s.t.} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \\
& \alpha_i \in \left[0, \frac{c}{n}\right] \; i = 1, \ldots, n.
\end{aligned}
$$

Kernels

Example

- Consider feature encoding for strings representing amino acids.
  - The characters are $\{A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$

Kernels

Example

- Consider feature encoding for strings representing amino acids.
  - The characters are $\boxed{\{A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V\}}$
- How should we relate the following strings…

```
IPTSALVKETLALLSTHRTLLIANETLRIPVPVHKNHQLCTEEIFQGIGTLESQTVQGGTV
ERLFKNLSLIKKYIDGQKKKCGEERRRVNQFLDYLQEFLGVMNTEWI
```

```
PHRRDLCSRSIWLARKIRSDLTALTESYVKHQGLWSELTEAERLQENLQAYRTFHVLLA
RLLEDQQVHFTPTEGDFHQAIHTLLLQVAAFAYQIEELMILLEYKIPRNEADGMLFEKK
LWGLKVLQELSQWTVRSIHDLRFISSHQTGIP
```

Kernels

Example

- Consider feature encoding for strings representing amino acids.
  - The characters are $\boxed{\{A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V\}}$
- How should we relate the following strings…

```
IPTSALVKETLALLSTHRTLLIANETLRIPVPVHKNHQLCTEEIFQGIGTLESQTVQGGTV
ERLFKNLSLIKKYIDGQKKKCGEERRRVNQFLDYLQEFLGVMNTEWI
```

$$\boxed{\kappa(x, x') = \sum_{s \in \mathcal{A}^*} w_s \phi_s(x) \phi_s(x')}$$

```
PHRRDLCSRSIWLARKIRSDLTALTESYVKHQGLWSELTEAERLQENLQAYRTFHVLLA
RLLEDQQVHFTPTEGDFHQAIHTLLLQVAAFAYQIEELMILLEYKIPRNEADGMLFEKK
LWGLKVLQELSQWTVRSIHDLRFISSHQTGIP
```

Summary

- Support Vector Machines
  - Hard Margin: Only applies to linearly separable data
  - Soft Margin: Allows for slack variables. Useful for outlier detection
- Rearranging Optimization Problems
  - Combine objective and constraint
  - Switch order of minimization / maximization
  - Lagrangians, First Order Conditions and Complementary Slackness
- Kernels
  - Replace features with relationships between features
  - We can use kernels for SVM because the problem depends on inner products – the linear kernel