EECS 445

Introduction to Machine Learning

SVM Dual and the Kernel Trick

Prof. Kutty

Review: Dual Formulation in General

Primal vs Dual formulation

objective function

constraints

original problem: $\min_{\bar{w}} f(\bar{w})$ s.t. $h_i(\bar{w}) \leq 0$ for $i = 1, \ldots, n$

primal

dual

Lagrangian: $L(\bar{w}, \bar{\alpha}) = f(\bar{w}) + \sum_{i=1}^{n} \alpha_i h_i(\bar{w})$ $\alpha_i \geq 0$

one corresponding to each constraints

Define: $g_p(\bar{w}) = \max_{\bar{\alpha}, \alpha_i \geq 0} L(\bar{w}, \bar{\alpha})$

$$g_p(\bar{w}) = \begin{cases} f(\bar{w}), & \text{if constraints are satisfied} \\ \infty, & \text{otherwise} \end{cases}$$

primal

$\min_{\bar{w}} g_p(\bar{w}) = \min_{\bar{w}} f(\bar{w})$ if constraints are satisfiable

Primal formulation $\min_{\bar{w}} \max_{\bar{\alpha}, \alpha_i \geq 0} L(\bar{w}, \bar{\alpha})$

Dual formulation $\max_{\bar{\alpha}, \alpha_i \geq 0} \min_{\bar{w}} L(\bar{w}, \bar{\alpha})$

Duality gap

Primal formulation $\min_{\bar{w}} \max_{\bar{\alpha}, \alpha_i \geq 0} L(\bar{w}, \bar{\alpha})$

Dual formulation $\max_{\bar{\alpha}, \alpha_i \geq 0} \min_{\bar{w}} L(\bar{w}, \bar{\alpha})$

1. The difference between these solutions is called the duality gap
2. The dual gives a lower bound on the solution of the primal
3. Under certain conditions*, however, the duality gap is zero

These conditions hold for our problem
i.e., the duality gap is zero

* quadratic convex objective, constraint functions affine, primal/dual feasible

Dual Formulation for SVMs

https://forms.gle/ffiBvNbPjHF8ghi77

Q1: what does $\bar{w}$ correspond to?

Q2: $h_i(\bar{w})$?

Q3: $L(\bar{w}, \bar{\alpha})$?

Support Vector Machines

Quadratic Program formulation

$\bar{x}^{(i)} \in \mathbb{R}^d$

$\bar{\theta} \in \mathbb{R}^d$

$\bar{\alpha} \in \mathbb{R}^n$

$1 - y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)}) \leq 0$

$$\min_{\bar{\theta}} \frac{\|\bar{\theta}\|^2}{2} \text{ subject to } y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)}) \geq 1 \text{ for } i \in \{1, \dots, n\}$$

original problem: $\min_{\bar{w}} f(\bar{w})$ s.t. $h_i(\bar{w}) \leq 0$ for $i = 1, \dots, n$

Lagrangian: $L(\bar{w}, \bar{\alpha}) = f(\bar{w}) + \sum_{i=1}^{n} \alpha_i h_i(\bar{w})$ $\alpha_i \geq 0$

1. Compose the Lagrangian

$$L(\bar{\theta}, \bar{\alpha}) = \frac{\|\bar{\theta}\|^2}{2} + \sum_{i=1}^{n} \alpha_i \left(1 - y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)})\right) \text{ with } \alpha_i \geq 0$$

2. Write the dual formulation

$$\max_{\bar{\alpha}, \alpha_i \geq 0} \min_{\bar{\theta}} L(\bar{\theta}, \bar{\alpha})$$

3. Rewrite in primal variable in terms of dual variables

Set $\nabla_{\bar{\theta}} L(\bar{\theta}, \bar{\alpha}) |_{\bar{\theta} = \bar{\theta}^*} = 0 \rightarrow \bar{\theta}^* = \sum_{i=1}^{n} \alpha_i y^{(i)} \bar{x}^{(i)}$

4. Simplify the dual formulation

Dual formulation

$$\max_{\bar{\alpha}, \alpha_i \geq 0} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \bar{x}^{(i)} \cdot \bar{x}^{(j)}$$

$$= \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y^{(i)} (\bar{\theta} \cdot \bar{x}^{(i)})$$

$$\nabla_{\bar{\theta}} \left(\frac{\|\bar{\theta}\|^2}{2} + \sum_{i=1}^{n} \alpha_i \left(1 - y^{(i)} (\bar{\theta} \cdot \bar{x}^{(i)})\right) \right) \Bigg|_{\bar{\theta} = \bar{\theta}^*} = 0$$

$$\bar{\theta}^* = \sum_{i=1}^{n} \alpha_i y^{(i)} \bar{x}^{(i)}$$

$$L(\bar{\theta}, \bar{\alpha}) =$$

$$\frac{\bar{\theta} \cdot \bar{\theta}}{2} + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y^{(i)} (\bar{\theta} \cdot \bar{x}^{(i)})$$

Substitute in $\bar{\theta}^*$

$$= \frac{1}{2} \left(\sum_{i=1}^{n} \alpha_i y^{(i)} \bar{x}^{(i)} \right) \cdot \left(\sum_{j=1}^{n} \alpha_j y^{(j)} \bar{x}^{(j)} \right)$$

$$+ \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y^{(i)} \left(\sum_{j=1}^{n} \alpha_j y^{(j)} \bar{x}^{(j)} \right) \cdot \bar{x}^{(i)}$$

$$= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \bar{x}^{(i)} \cdot \bar{x}^{(j)}$$

$$+ \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \bar{x}^{(i)} \cdot \bar{x}^{(j)}$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \bar{x}^{(i)} \cdot \bar{x}^{(j)}$$

Dual variables and Support Vectors

$$\max_{\bar{\alpha}, \alpha_i \geq 0} \min_{\bar{\theta}} L(\bar{\theta}, \bar{\alpha})$$

constraints

$y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)}) \geq 1$

where $L(\bar{\theta}, \bar{\alpha}) = \frac{||\bar{\theta}||^2}{2} + \sum_{i=1}^{n} \alpha_i (1 - y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)}))$

Let optimal values be given by $\bar{\theta}^*$ and $\hat{\alpha}_1, \ldots, \hat{\alpha}_n$

Solution satisfies "complementary slackness constraints":

$$\hat{\alpha}_i > 0 \rightarrow y^{(i)} \bar{\theta}^* \cdot \bar{x}^{(i)} = 1$$ (support vector)

$$\hat{\alpha}_i = 0 \leftarrow y^{(i)} \bar{\theta}^* \cdot \bar{x}^{(i)} > 1$$ (non-support vector)

In other words, either the primal inequality is satisfied with equality or the dual variable is zero.

Dual variables and Support Vectors

- for hard margin SVMs support vectors can include *only*
  - points on the margin

- for soft margin SVMs support vectors can include *only*
  - points on the margin
  - points on the “wrong side” of the margin
    - misclassified points
    - points within the margin

Dual variables and Support Vectors

$$\max_{\bar{\alpha}, \alpha_i \geq 0} \min_{\bar{\theta}} L(\bar{\theta}, \bar{\alpha})$$

constraints

$$y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)}) \geq 1$$

where $L(\bar{\theta}, \bar{\alpha}) = \frac{||\bar{\theta}||^2}{2} + \sum_{i=1}^{n} \alpha_i (1 - y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)}))$

Let optimal values be given by $\bar{\theta}^*$ and $\hat{\alpha}_1, \ldots, \hat{\alpha}_n$

$$\boxed{\bar{\theta}^* = \sum_{i=1}^{n} \hat{\alpha}_i y^{(i)} \bar{x}^{(i)}}$$

Support vectors are the most important datapoints in the dataset → *non-zero duals* → separating hyperplane depends on these

Dual SVM with Offset

The Primal (with offset):

$$\min_{\bar{\theta},\ b} \frac{1}{2}||\bar{\theta}||^2 \text{ subject to } y^{(i)}(\bar{\theta}\cdot\bar{x}^{(i)} + \boxed{b}) \geq 1 \text{ for } i = 1, ..., n$$

The Dual (with offset):

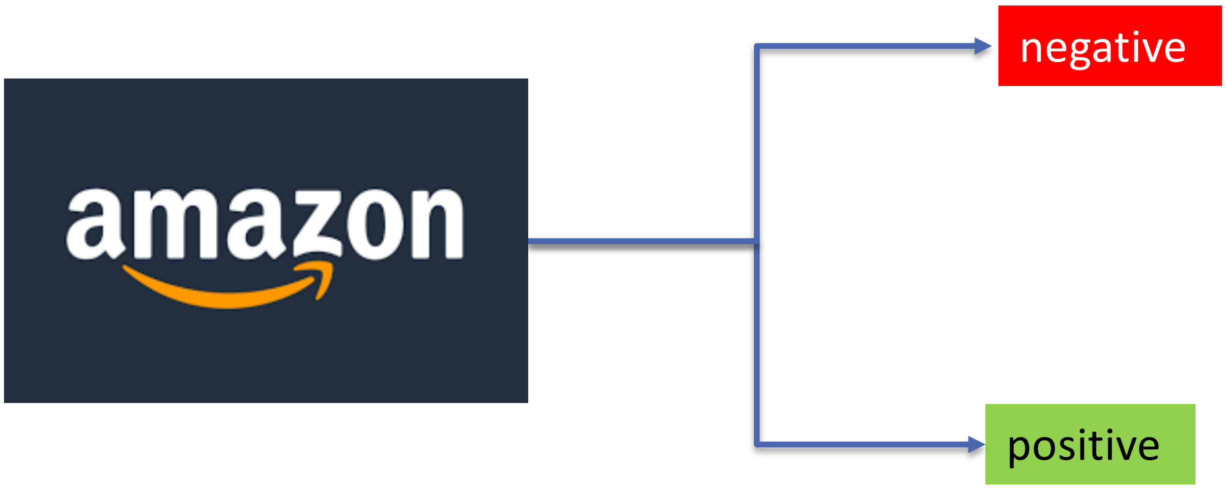
$$\max_{\bar{\alpha},\alpha_i\geq 0} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y^{(i)} y^{(j)} \bar{x}^{(i)}\cdot\bar{x}^{(j)}$$

$$\boxed{\text{subject to } \sum_{i=1}^{n} \alpha_i y^{(i)} = 0}$$

Additional constraint

For derivation see discussion notes

SVMs applied to text data



Project 1: Overview

- **Section 1: Introduction**
  - read this! It provides context and includes details on packages used etc.
- **Section 2: Feature Extraction**
  - (n, d) feature matrix, where each row represents a review, and each column represents whether or not a specific word appeared in that review.
- **Section 3: Hyperparameter and Model Selection**
  - learn a classifier to classify the training data into positive and negative labels using different models
- **Section 4: Asymmetric Cost Functions and Class Imbalance**
  - assigning different weights to slack variables for -vely and +vely labeled datapoints
- **Section 5: Identifying Bias**
  - gender bias in word embeddings
- **Section 6: Challenge**
  - goal is to learn a multiclass classifier using the SVC or LinearSVC class to predict the true ratings of the *held-out test set*.
  - we will evaluate your performance based on:
    - effort (present your analysis)
    - accuracy in the context of the performance of the class at-large
- **Section 7: Code Appendix**

See Appendix for additional ideas

Project 1 Submission procedure

This spec contains 20 pages, including Appendices with approximate run-times for programming problems, as well as a list of topics, concepts, and further reading. + IA/GSI OS

You will submit the project components to three separate Gradescope assignments:

- Submit your write-up for sections 2-6 to the Gradescope assignment titled "Project 1 Writeup"
- Submit your file `uniqname.csv` containing the label predictions for the heldout data to the Gradescope assignment titled "Project 1 Challenge Submission".
- Submit all of your project code to the Gradescope assignment titled "Project 1 Code Appendix". You should include any code from `project1.py` and `helper.py`, as well as any additional functions or code you wrote to generate the output you reported in your write-up. You can submit your code as is, including any comments or print statements. Accepted file formats for the Gradescope submission include py files, zip files, ipynb files, and PDFs of your code. Your code appendix will be manually graded for effort and completeness.

Project 1 Quickstart will be held tomorrow! See calendar for details.
Slides and notes will be provided on canvas. Recording available within ~24 hours.

SVMs and the Kernel Trick

https://forms.gle/ffiBvNbPjHF8ghi77

Given $S_n = \{\bar{x}^{(i)}, y^{(i)}\}_{i=1}^{n}$

Learn maximum margin classifier $sign(\bar{\theta}^* \cdot \bar{x})$

$$\boxed{\min_{\bar{\theta}} \frac{\|\bar{\theta}\|^2}{2} \text{ subject to } y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)}) \geq 1 \text{ for } i \in \{1, \ldots, n\}}$$

Goal: rewrite in dual form

$$\max_{\bar{\alpha}, \alpha_i \geq 0} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \bar{x}^{(i)} \cdot \bar{x}^{(j)}$$

Output of this optimization problem is $\hat{\bar{\alpha}}$:

$$\bar{\theta}^* = \sum_{i=1}^{n} \hat{\alpha}_i y^{(i)} \bar{x}^{(i)}$$

Why do this?

Given $S_n = \{\bar{x}^{(i)}, y^{(i)}\}_{i=1}^{n}$.

Suppose we wanted to map to a higher dimensional space.

The usual way $S_n = \{\phi(\bar{x}^{(i)}), y^{(i)}\}_{i=1}^{n}$

Solve dual form

replace with $K(\bar{x}^{(i)}, \bar{x}^{(j)})$

$$\max_{\bar{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \left(\phi(\bar{x}^{(i)}) \cdot \phi(\bar{x}^{(j)})\right)$$

$$\text{subject to } \alpha_i \geq 0 \quad \forall i = 1, .., n$$

Sometimes… $\phi(\bar{x}^{(i)}) \cdot \phi(\bar{x}^{(j)})$ can be computed much more efficiently than separately computing $\phi(\bar{x}^{(i)})$ and $\phi(\bar{x}^{(j)})$

$$\begin{array}{ccc} \bar{x}^{(i)}, \bar{x}^{(j)} & \xrightarrow{\text{some computation in } d\text{-dimensional space}} & \text{scalar} \\ \downarrow & & \uparrow \\ \phi(\bar{x}^{(i)}) \cdot \phi(\bar{x}^{(j)}) & \longrightarrow & \end{array}$$

Kernels and Feature Maps

Feature map

ϕ takes input $\bar{x} \in \mathcal{X}$ (e.g., $\bar{x} \in \mathbb{R}^d$) and maps it to feature space $\mathcal{F}$ (e.g., $\mathbb{R}^p$)

Each kernel has an associated feature mapping

$$K(\bar{u}, \bar{v}) = \phi(\bar{u}) \cdot \phi(\bar{v})$$

$$\phi : \mathcal{X} \rightarrow \mathcal{F}$$
$$K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$$

Intuitively, the kernel function takes two inputs and gives their similarity in the feature space

Classifying a new example $\bar{x}$

Previously: $h(\bar{x}) = sign(\bar{\theta} \cdot \bar{x})$

assume α_i represent optimal values

Recall: $$\bar{\theta}^* = \sum_{i=1}^{n} \alpha_i y^{(i)} \bar{x}^{(i)}$$

So: $$h(\bar{x}) = sign\left(\left(\sum_{i=1}^{n} \alpha_i y^{(i)} \bar{x}^{(i)}\right) \cdot \bar{x}\right)$$

Now: $$\bar{\theta}^* = \sum_{i=1}^{n} \alpha_i y^{(i)} \phi(\bar{x}^{(i)})$$

$$= sign\left(\sum_{i=1}^{n} \alpha_i y^{(i)} \boxed{\phi(\bar{x}^{(i)}) \cdot \phi(\bar{x})}\right)$$

So: $$h(\bar{x}) = sign\left(\left(\sum_{i=1}^{n} \alpha_i y^{(i)} \phi(\bar{x}^{(i)})\right) \cdot \phi(\bar{x})\right)$$

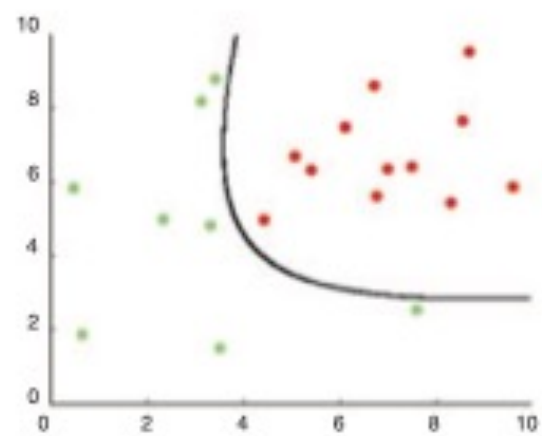
$$= sign\left(\sum_{i=1}^{n} \alpha_i y^{(i)} K(\bar{x}^{(i)}, \bar{x})\right)$$

Quadratic Decision Boundary

Feature Map $\phi(\bar{u}) = \left[u_1^2, u_2^2, \sqrt{2}u_1 u_2\right]^{\mathrm{T}}$ for $\bar{u} \in \mathbb{R}^2$.

Here $\phi: \mathbb{R}^2 \rightarrow \mathbb{R}^3$

$\bar{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$



Kernels and Feature Maps: example

Consider a Feature Map $\phi(\bar{u}) = \left[u_1^2, u_2^2, \sqrt{2}u_1 u_2\right]^{\mathrm{T}}$ for $\bar{u} \in \mathbb{R}^2$.

Here $\phi: \mathbb{R}^2 \to \mathbb{R}^3$

$$\phi(\bar{u}) \cdot \phi(\bar{v}) = \left[u_1^2, u_2^2, \sqrt{2}u_1 u_2\right]^{\mathrm{T}} \cdot \left[v_1^2, v_2^2, \sqrt{2}v_1 v_2\right]^{\mathrm{T}}$$

$$= u_1^2 v_1^2 + u_2^2 v_2^2 + 2u_1 u_2 v_1 v_2 = (\bar{u} \cdot \bar{v})^2$$

Kernel $K(\bar{u}, \bar{v}) = (\bar{u} \cdot \bar{v})^2 = \phi(\bar{u}) \cdot \phi(\bar{v})$

where

Feature Map $\phi(\bar{u}) = \left[u_1^2, u_2^2, \sqrt{2}u_1 u_2\right]^{\mathrm{T}}$

$\bar{u} \cdot \bar{v}$

$= \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = u_1 v_1 + u_2 v_2$

Kernel Algebra

Let K1 and K2 be valid kernels, then the following are valid kernels:

$$K(\bar{x}, \bar{z}) = K_1(\bar{x}, \bar{z}) + K_2(\bar{x}, \bar{z})$$ sum

$$K(\bar{x}, \bar{z}) = \alpha K_1(\bar{x}, \bar{z})$$ scalar product $\alpha > 0$

$$K(\bar{x}, \bar{z}) = K_1(\bar{x}, \bar{z}) K_2(\bar{x}, \bar{z})$$ direct product

Examples of Valid Kernels

Linear Kernel

$$K(\bar{u}, \bar{v}) = \bar{u} \cdot \bar{v}$$

Quadratic Kernel

$$K(\bar{u}, \bar{v}) = (\bar{u} \cdot \bar{v} + r)^2 \text{ with } r \geq 0$$

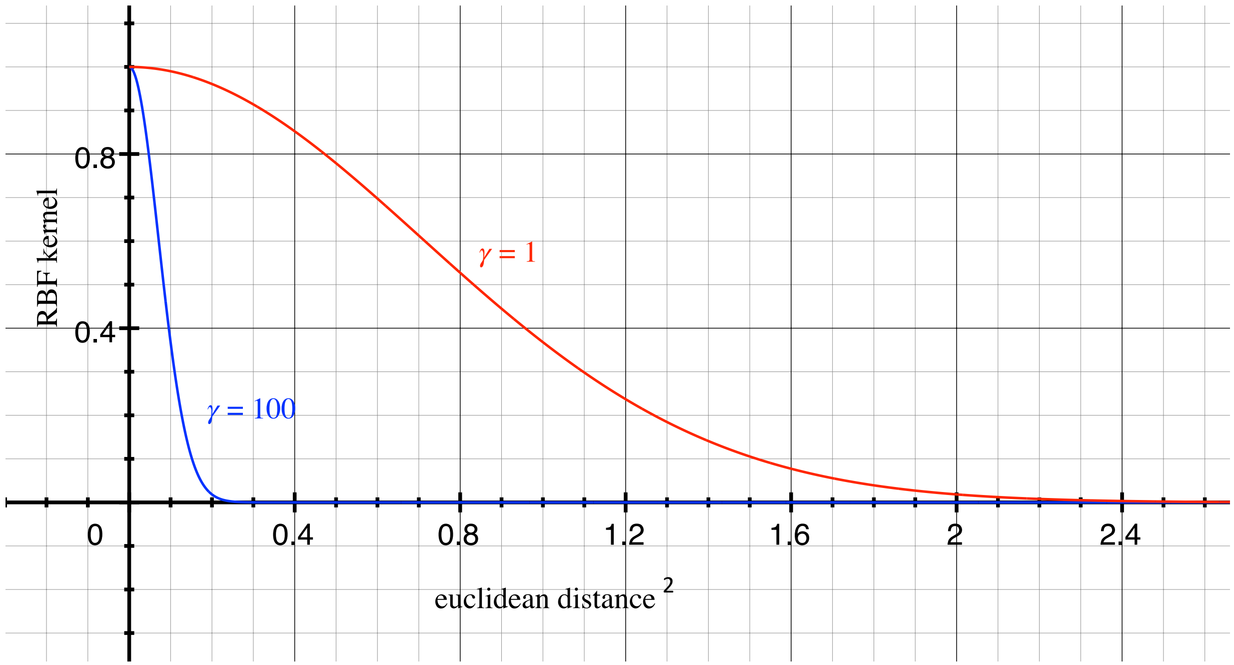
Radial Basis Function

RBF Kernel (aka Gaussian Kernel)

$$K(\bar{u}, \bar{v}) = \exp(-\gamma \|\bar{u} - \bar{v}\|^2) \quad \text{with } \gamma \geq 0$$

hyperparameter

Gaussian Kernel



RBF Kernel (aka Gaussian Kernel)

$$K(\bar{u}, \bar{v}) = \exp(-\gamma \|\bar{u} - \bar{v}\|^2)$$

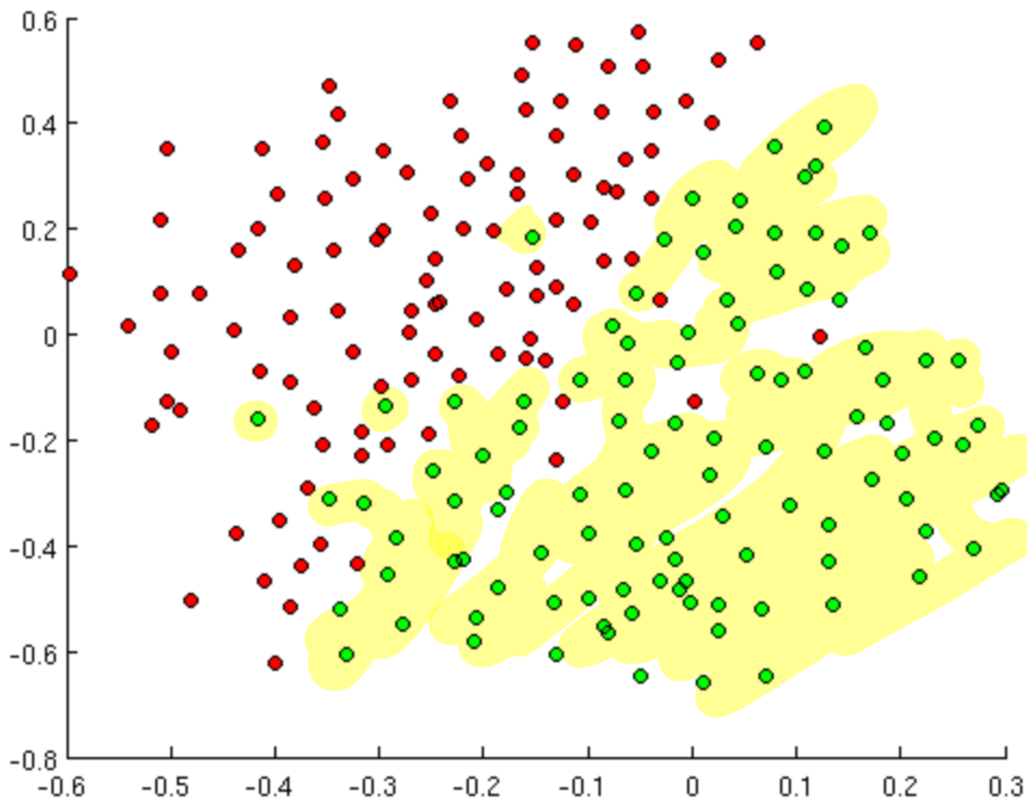
$\bar{u} = \bar{v}$

$0 \rightarrow e^0 = 1$

Visualization

RBF kernel

$$K(\bar{x}^{(i)}, \bar{x}^{(j)}) = exp(-\gamma||\bar{x}^{(i)} - \bar{x}^{(j)}||^2)$$

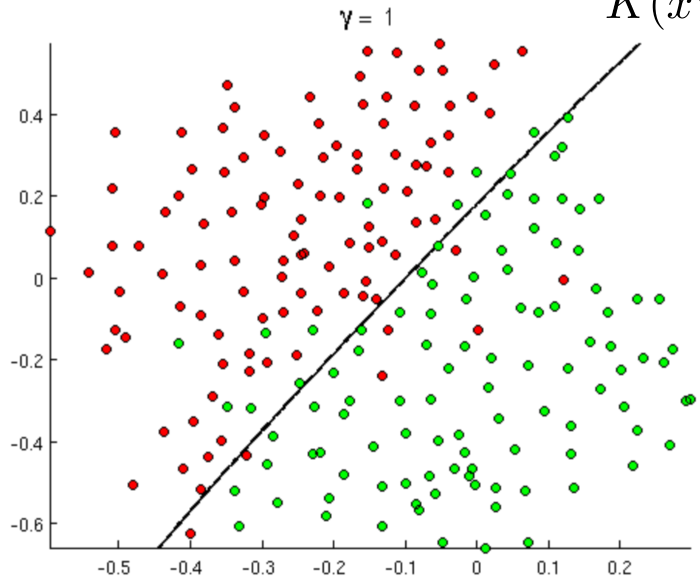


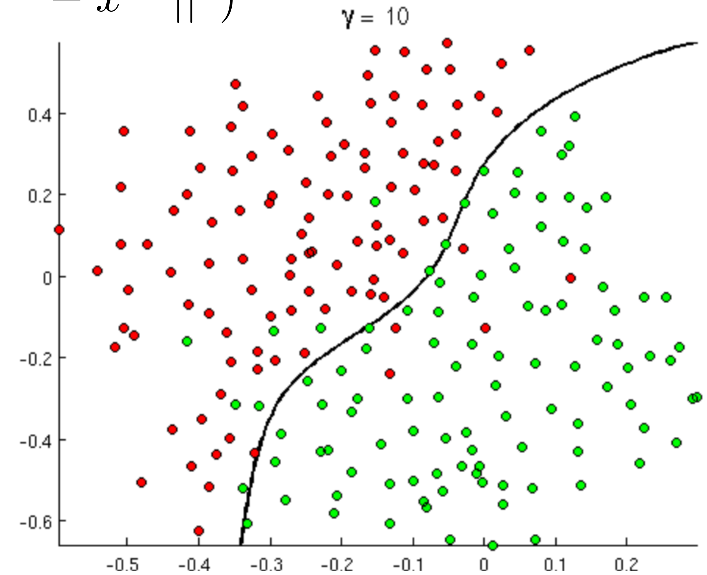
This example courtesy Andrew Ng

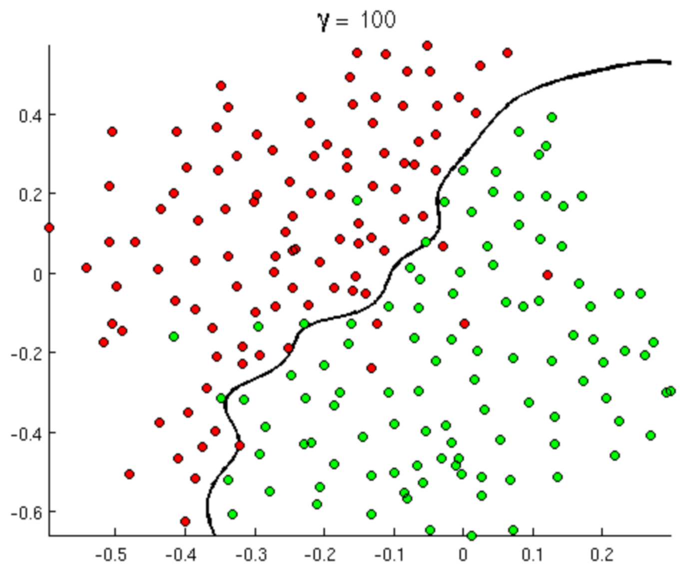
Visualization

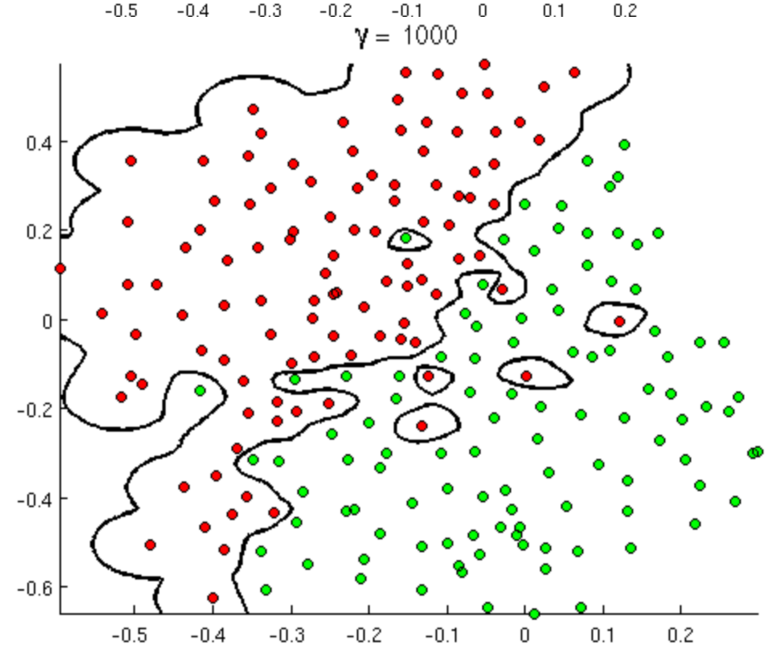
SVM with RBF kernel

$$K(\bar{x}^{(i)}, \bar{x}^{(j)}) = exp(-\gamma||\bar{x}^{(i)} - \bar{x}^{(j)}||^2)$$

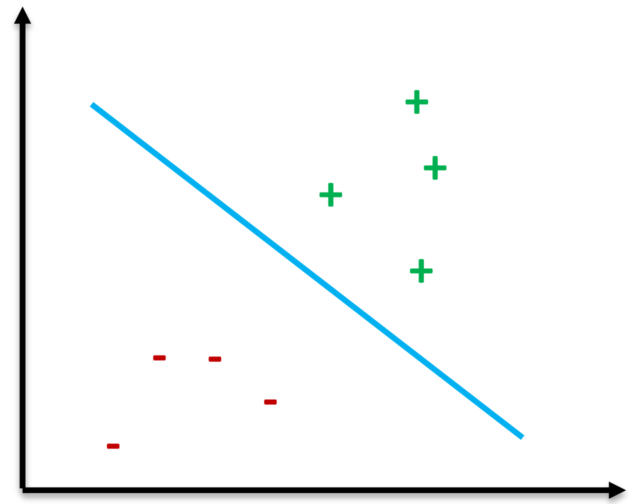








Interpretability of Linear decision boundaries



$\text{sign}(\theta_1 x_1 + \theta_2 x_2)$

what does $\bar{\theta}$ mean?

Interpretability and Kernels

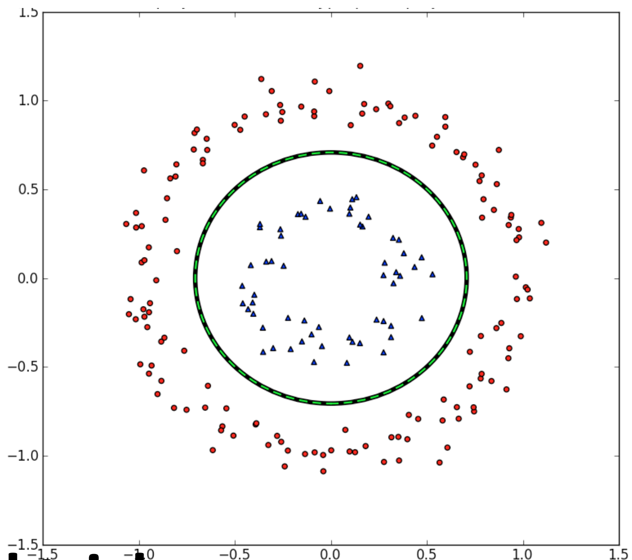
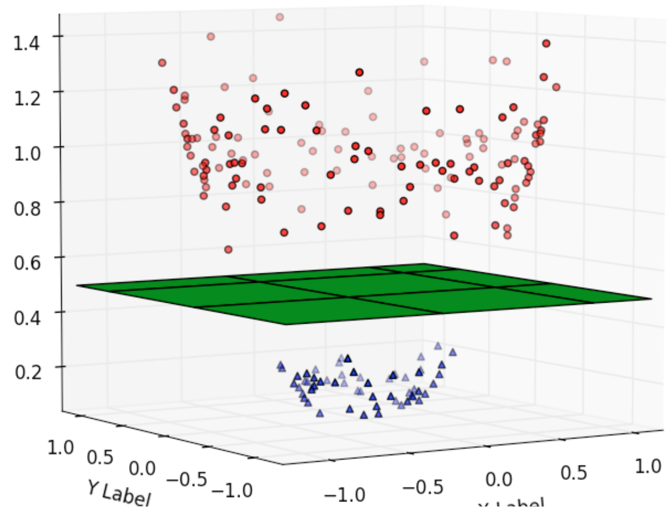


image source: http://www.eric-kim.net



Kernel trick

maps data to a higher dim. space in which there exists a separating hyperplane (corresponds to a non-linear decision boundary in the original space)

- never have to explicitly compute feature mappings
- RBF kernel maps data to an infinite dimensional feature space

Problem: need not recover model parameters, classifier becomes a **black box**

RBF Kernels

$$K(\bar{u},\bar{v}) = \exp(-\gamma\|\bar{u}-\bar{v}\|^2)$$

$$K(\bar{u},\bar{v}) = \exp(-\gamma\|\bar{u}-\bar{v}\|^2)$$

$\bar{u} \in \mathbb{R}^2$

$\bar{v} \in \mathbb{R}^2$

$$\|\bar{u}-\bar{v}\|^2 = \|(u_1 \quad u_2)^T - (v_1 \quad v_2)^T\|^2$$

$$= \left\| \begin{bmatrix} u_1 - v_1 \\ u_2 - v_2 \end{bmatrix} \right\|^2$$

$$= (u_1 - v_1)^2 + (u_2 - v_2)^2$$

$$= u_1^2 + v_1^2 - 2u_1v_1 + u_2^2 + v_2^2 - 2u_2v_2$$

$$= \|\bar{u}\|^2 + \|\bar{v}\|^2 - 2(\bar{u}\cdot\bar{v})$$

$$K(\bar{u},\bar{v}) = \exp(-\gamma\|\bar{u}\|^2 - \gamma\|\bar{v}\|^2 + 2\gamma(\bar{u}\cdot\bar{v}))$$

$$= \exp(-\gamma\|\bar{u}\|^2)\exp(-\gamma\|\bar{v}\|^2)\exp(2\gamma(\bar{u}\cdot\bar{v}))$$

Recall Taylor series expansion

$$e^{x} = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

Idea $$e^{\bar{u}\cdot\bar{v}} = \frac{(\bar{u}\cdot\bar{v})^0}{0!} + \frac{(\bar{u}\cdot\bar{v})^1}{1!} + \frac{(\bar{u}\cdot\bar{v})^2}{2!} + \cdots$$

Sum of polynomial kernels

$\rightarrow$ feature map of the RBF kernel is infinite dimensional