UVA CS 4501 - 001 / 6501 – 007 Introduction to Machine Learning and Data Mining

Lecture 11: Classification with Support Vector Machine (Review + Practical Guide)

Yanjun Qi / Jane

University of Virginia
Department of
Computer Science

Where are we ? ➔ Five major sections of this course

- ❑ Regression (supervised)
- ❑ Classification (supervised)
- ❑ Unsupervised models
- ❑ Learning theory
- ❑ Graphical models

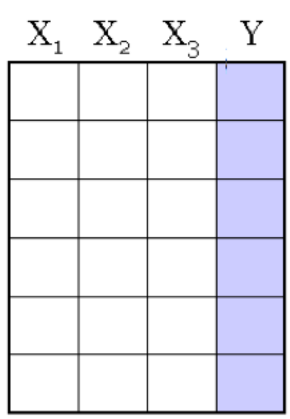
Where are we ? ➔ Three major sections for classification

- We can divide the large variety of classification approaches into roughly three major types

1. Discriminative
   - directly estimate a decision rule/boundary
   - e.g., support vector machine, decision tree

2. Generative:
   - build a generative statistical model
   - e.g., naïve bayes classifier, Bayesian networks

3. Instance based classifiers
   - Use observation directly (no models)
   - e.g. K nearest neighbors



A Dataset for binary classification

$f : \boxed{X} \longrightarrow \boxed{Y}$

Output as Binary Class Label: 1 or -1

- **Data**/*points/instances/examples/samples/records*: [rows]
- **Features**/*attributes/dimensions/independent variables/covariates/ predictors/regressors*: [columns, except the last]
- **Target**/*outcome/response/label/dependent variable*: special column to be predicted [last column]

Today: Review & Practical Guide

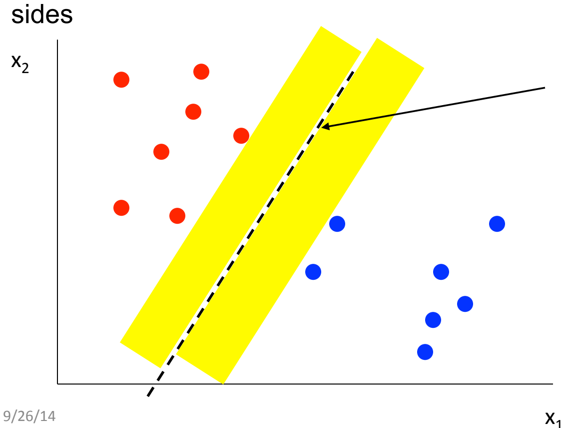
- ❑ Support Vector Machine (SVM)

review

- ✓ Large Margin Linear Classifier
- ✓ Define Margin (M) in terms of model parameter
- ✓ Optimization to learn model parameters (w, b)
- ✓ Non linearly separable case
- ✓ Optimization with dual form
- ✓ Nonlinear decision boundary
- ✓ Practical Guide

Max margin classifiers

- Instead of fitting all points, focus on boundary points
- Learn a boundary that leads to the largest margin from points on both sides



Why?

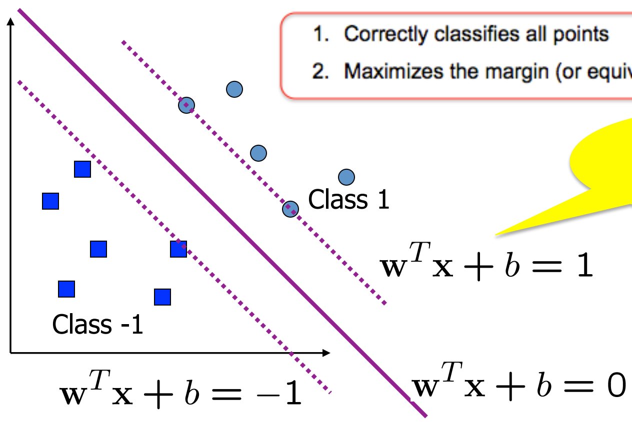
- Intuitive, 'makes sense'
- Some theoretical support
- Works well in practice

When linearly Separable Case

- The decision boundary should be as far away from the data of both classes as possible

1. Correctly classifies all points
2. Maximizes the margin (or equivalently minimizes $w^T w$)

W is a p-dim vector; b is a scalar

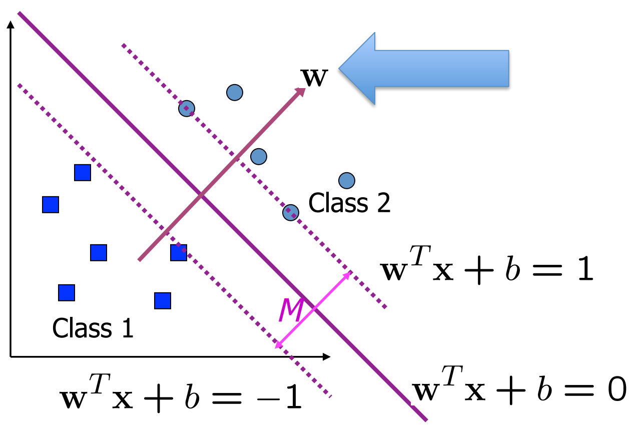


Today: Review & Practical Guide

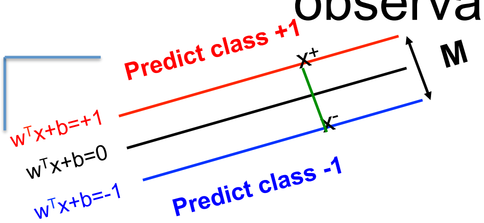
- ❑ Support Vector Machine (SVM)
  - ✓ Large Margin Linear Classifier
  - review ✓ Define Margin (M) in terms of model parameter
  - ✓ Optimization to learn model parameters (w, b)
  - ✓ Non linearly separable case
  - ✓ Optimization with dual form
  - ✓ Nonlinear decision boundary
  - ✓ Practical Guide

Maximizing the margin: observation-1

- **Observation 1: the vector w is orthogonal to the +1 plane**



Maximizing the margin: observation-2



Classify as +1 if $w^Tx+b \geq 1$

Classify as -1 if $w^Tx+b \leq -1$

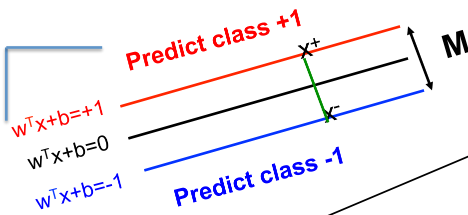
Undefined if $-1 < w^Tx+b < 1$

- Observation 1: the vector w is orthogonal to the +1 and -1 planes
- Observation 2: if x^+ is a point on the +1 plane and x^- is the closest point to x^+ on the -1 plane then

$$x^+ = \lambda w + x^-$$

Since w is orthogonal to both planes we need to 'travel' some distance along w to get from x^+ to x^-

Putting it together



- $w^T x^+ + b = +1$
- $w^T x^- + b = -1$
- $x^+ = \lambda w + x^-$
- $| x^+ - x^- | = M$

$w^T x^+ + b = +1$

$\Rightarrow$

$w^T (\lambda w + x^-) + b = +1$

$\Rightarrow$

$w^Tx^- + b \; + \lambda w^Tw = +1$

$\Rightarrow$

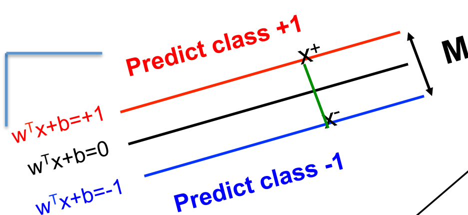
$-1 \; + \lambda w^Tw = +1$

$\Rightarrow$

$\lambda = 2/w^Tw$

We can now define M in terms of w and b

Putting it together



- $w^T x^+ + b = +1$
- $w^T x^- + b = -1$
- $x^+ = \lambda w + x^-$
- $| x^+ - x^- | = M$
- $\lambda = 2/w^Tw$

$M = |x^+ - x^-|$

$\Rightarrow$

$$M = |\lambda w| = \lambda |w| = \lambda\sqrt{w^Tw}$$

$\Rightarrow$

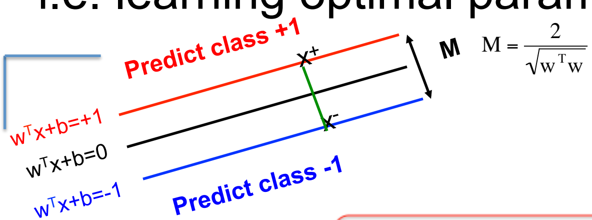
$$M = 2\frac{\sqrt{w^Tw}}{w^Tw} = \frac{2}{\sqrt{w^Tw}}$$

We can now define M in terms of w and b

Today: Review & Practical Guide

- ❑ Support Vector Machine (SVM)
  - ✓ Large Margin Linear Classifier
  - ✓ Define Margin (M) in terms of model parameter
  - review ✓ Optimization to learn model parameters (w, b)
  - ✓ Non linearly separable case
  - ✓ Optimization with dual form
  - ✓ Nonlinear decision boundary
  - ✓ Practical Guide

Optimization Step
i.e. learning optimal parameter for SVM



1. Correctly classifies all points
2. Maximizes the margin (or equivalently minimizes w^Tw)

Min $(w^Tw)/2$

subject to the following constraints:

For all x in class + 1

$w^Tx+b \geq 1$

For all x in class - 1

$w^Tx+b \leq -1$

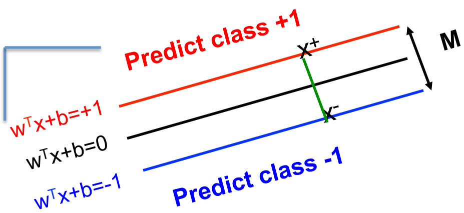
A total of n constraints if we have n input samples

$$\underset{\mathbf{w},b}{\operatorname{argmin}} \sum_{i=1}^{p} w_i^2$$

subject to $\forall \mathbf{x}_i \in Dtrain : y_i(\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1$

SVM as a QP problem

R as I matrix, d as zero vector, c as 0 value



$$M = \frac{2}{\sqrt{w^T w}}$$

$$\min_U \frac{u^T R u}{2} + d^T u + c$$

subject to n inequality constraints:

$$a_{11}u_1 + a_{12}u_2 + \ldots \le b_1$$
$$\vdots \quad \vdots \quad \vdots$$
$$a_{n1}u_1 + a_{n2}u_2 + \ldots \le b_n$$

and k equivalency constraints:

$$a_{n+1,1}u_1 + a_{n+1,2}u_2 + \ldots = b_{n+1}$$
$$\vdots \quad \vdots \quad \vdots$$
$$a_{n+k,1}u_1 + a_{n+k,2}u_2 + \ldots = b_{n+k}$$

Min $(w^T w)/2$

subject to the following inequality constraints:

For all x in class + 1

$w^T x + b \ge 1$

For all x in class - 1

$w^T x + b \le -1$

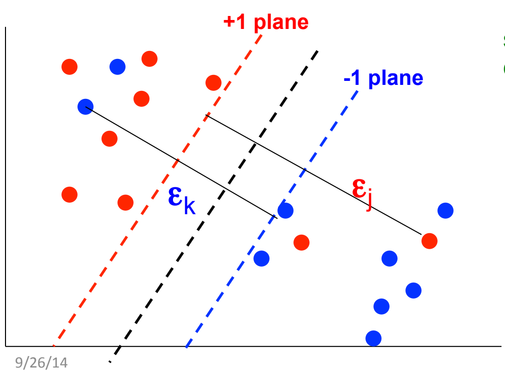
A total of n constraints if we have n input samples

Today: Review & Practical Guide

- ❑ Support Vector Machine (SVM)
  - ✓ Large Margin Linear Classifier
  - ✓ Define Margin (M) in terms of model parameter
  - ✓ Optimization to learn model parameters (w, b)
  - review ✓ Non linearly separable case
  - ✓ Optimization with dual form
  - ✓ Nonlinear decision boundary
  - ✓ Practical Guide

Non linearly separable case

- Instead of minimizing the number of misclassified points we can minimize the (relative) ***distance*** between these points and their correct plane



The new optimization problem is:

$$\min_w \frac{w^T w}{2} + \sum_{i=1}^{n} C\varepsilon_i$$

subject to the following inequality constraints:

For all x_i in class + 1

$w^T x+b \geq 1- \varepsilon_i$

For all x_i in class - 1

$w^T x+b \leq -1+ \varepsilon_i$

A total of n constraints

For all i

$\varepsilon_i \geq 0$

Another n constraints

Where we are

Two optimization problems: For the separable and non separable cases

$$\min_w \frac{w^T w}{2}$$

For all x in class + 1

$w^T x+b \geq 1$

For all x in class - 1

$w^T x+b \leq -1$

$$\min_w \frac{w^T w}{2} + \sum_{i=1}^{n} C\varepsilon_i$$

For all x_i in class + 1

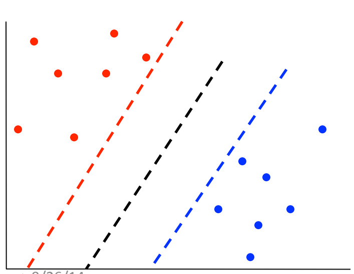
$w^T x+b \geq 1- \varepsilon_i$

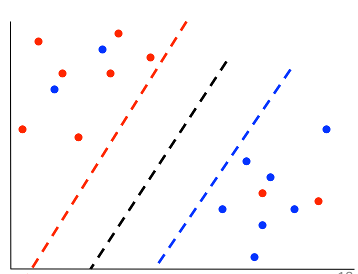
For all x_i in class - 1

$w^T x+b \leq -1+ \varepsilon_i$

For all i

$\varepsilon_i \geq 0$





Today: Review & Practical Guide

- Support Vector Machine (SVM)
  - ✓ Large Margin Linear Classifier
  - ✓ Define Margin (M) in terms of model parameter
  - ✓ Optimization to learn model parameters (w, b)
  - ✓ Non linearly separable case
  - review ✓ Optimization with dual form
  - ✓ Nonlinear decision boundary
  - ✓ Practical Guide

Where we are

Two optimization problems: For the separable and non separable cases

Min $(w^T w)/2$

For all x in class + 1

$w^T x+b \geq 1$

For all x in class - 1

$w^T x+b \leq -1$

$$\min_w \frac{w^T w}{2} + \sum_{i=1}^{n} C\varepsilon_i$$

For all x_i in class + 1

$w^T x+b \geq 1- \varepsilon_i$

For all x_i in class - 1

$w^T x+b \leq -1+ \varepsilon_i$

For all i

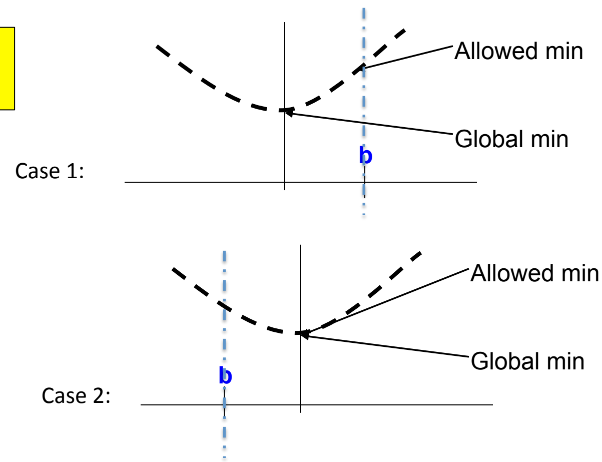
$\varepsilon_i \geq 0$

- Instead of solving these QPs directly we will solve a dual formulation of the SVM optimization problem
- The main reason for switching to this type of representation is that it would allow us to use a neat trick that will make our lives easier (and the run time faster)

Optimization Review:
Constrained Optimization

$\min_u u^2$

s.t. $u \geq b$



Optimization Review:
Constrained Optimization with Lagrange

- When equal constraints
- ➔ optimize $f(x)$, subject to $g_i(x)=0$

- Method of Lagrange multipliers: convert to a higher-dimensional problem
- Minimize

$$f(x) + \sum \lambda_i g_i(x)$$

- w.r.t. $(x_1 \ldots x_n ; \lambda_1 \ldots \lambda_k)$

Introducing a Lagrange multiplier for each constraint
Construct the Lagrangian for the original optimization problem

Optimization Review: Dual Problem

- Using dual problem
  - Constrained optimization → unconstrained optimization
- Need to change maximization to minimization
- Only valid when the original optimization problem is convex/concave (strong duality)

Dual Problem

$$\lambda^* = \arg\min_{\lambda} l(\lambda)$$

Primal Problem

$$x^* = \arg\max_{x} f(x)$$

$$\text{subject to } g(x) = c$$

$x^*=\lambda^*$
When convex/concave

$$l(\lambda) = \sup_{x}(f(x) + \lambda(g(x) - c))$$

An alternative (dual) representation for SVM QP

Here α is the lagrange multiplier variable

Min $(w^T w)/2$

$(w^T x_i + b) y_i \geq 1$

- We will start with the linearly separable case
- Instead of encoding the correct classification rule a constraint we will use Lagrange multiplies to encode it as part of the our minimization problem

Recall that Lagrange multipliers can be applied to turn the following problem:

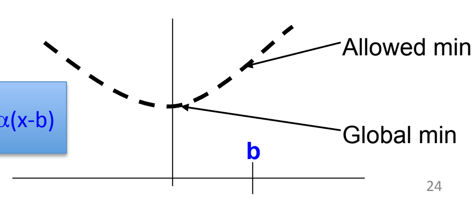
$\min_x x^2$

s.t. $x \geq b$ → $b-x \leq 0$

To

$\text{Min}_{x,\alpha}\, x^2 + \alpha(b-x)$ ← $\min_x \max_\alpha x^2 - \alpha(x-b)$

s.t. $\alpha \geq 0$



Lagrange multiplier for SVMs / Linearly Separable Case

Dual formulation

$$\min_{w,b} \max_{\alpha} \frac{\mathrm{w}^T\mathrm{w}}{2} - \sum_i \alpha_i[(\mathrm{w}^T x_i + b)y_i - 1]$$

$$\alpha_i \geq 0 \qquad \forall i$$

Original formulation

Min $(w^Tw)/2$

$(w^Tx_i+b)y_i \geq 1$

Using this new formulation we can derive w and b by taking the derivative w.r.t. w and α leading to:

$$w = \sum_i \alpha_i x_i y_i$$

$$b = y_i - \mathrm{w}^T x_i$$

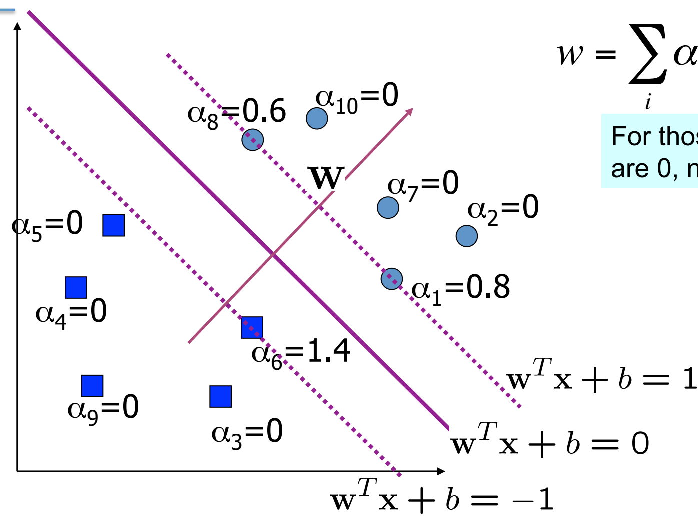
$$for \quad i \quad s.t. \quad \alpha_i > 0$$

Set partial derivatives to 0

Finally, taking the derivative w.r.t. b we get:

$$\sum_i \alpha_i y_i = 0$$

A Geometrical Interpretation



$$w = \sum_i \alpha_i x_i y_i$$

For those α_i that are 0, no influence

Dual SVM for linearly separable case

Substituting w into our target function and using the additional constraint we get:

$$\min_{w,b} \frac{\mathrm{w}^T\mathrm{w}}{2} - \sum_i \alpha_i [(\mathrm{w}^T x_i + b) y_i - 1]$$

$$\alpha_i \geq 0 \qquad \forall i$$

$$w = \sum_i \alpha_i x_i y_i$$

$$b = y_i - \mathrm{w}^T x_i$$

$$for \quad i \quad s.t. \quad \alpha_i > 0$$

$$\sum_i \alpha_i y_i = 0$$

Dual formulation

$$\max_\alpha \sum_i \alpha_i - \frac{1}{2} \sum_{\mathrm{i,j}} \alpha_i \alpha_j \mathrm{y_i} y_j \mathbf{x_i}^T \mathbf{x_j}$$

$$\sum_i \alpha_i \mathrm{y_i} = 0$$

$$\alpha_i \geq 0 \qquad \forall i$$

Easier than original QP, a QP solver can be used to find α_i

Dual SVM for linearly separable case

Our dual target function:

$$\max_\alpha \sum_i \alpha_i - \frac{1}{2} \sum_{\mathrm{i,j}} \alpha_i \alpha_j \mathrm{y_i} y_j \mathbf{x_i}^T \mathbf{x_j}$$

$$\sum_i \alpha_i \mathrm{y_i} = 0$$

$$\alpha_i \geq 0 \qquad \forall i$$

Dot product among all training samples

Dot product of test sample with all training samples

To evaluate a new sample x_j we need to compute:

$$\mathrm{w}^T x_j + b = \sum_{\mathrm{i}} \alpha_i \mathrm{y_i} \mathbf{x_i}^T \mathbf{x_j} + b$$

Dual formulation for non linearly separable case

Dual target function:

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

$$\sum_i \alpha_i y_i = 0$$

$$C > \alpha_i \geq 0, \forall i$$

Hyperparameter C should be tuned through k-folds CV

The only difference is that the α_i's are now bounded

To evaluate a new sample x_j we need to compute:

$$w^T x_j + b = \sum_i \alpha_i y_i \mathbf{x}_i^T \mathbf{x}_j + b$$

This is very similar to the optimization problem in the linear separable case, except that there is an upper bound C on α_i now

Once again, a QP solver can be used to find α_i

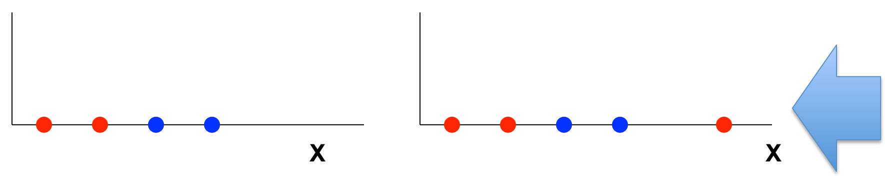
Today: Review & Practical Guide

- Support Vector Machine (SVM)
  - ✓ Large Margin Linear Classifier
  - ✓ Define Margin (M) in terms of model parameter
  - ✓ Optimization to learn model parameters (w, b)
  - ✓ Non linearly separable case
  - ✓ Optimization with dual form
  - review ✓ Nonlinear decision boundary
  - ✓ Practical Guide

Classifying in 1-d

Can an SVM correctly classify this data?

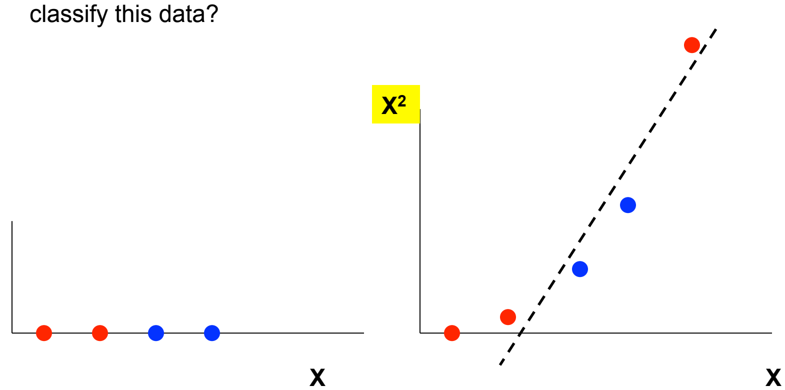
What about this?



Classifying in 1-d

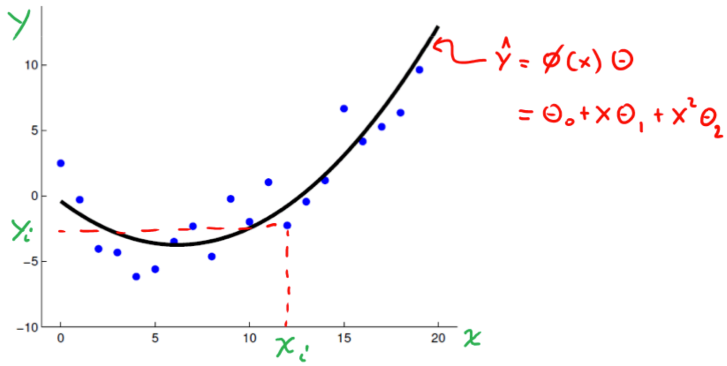
Can an SVM correctly classify this data?

And now? (extend with polynomial basis)



RECAP: **Polynomial regression**

For example, $\phi(x) = [1, x, x^2]$

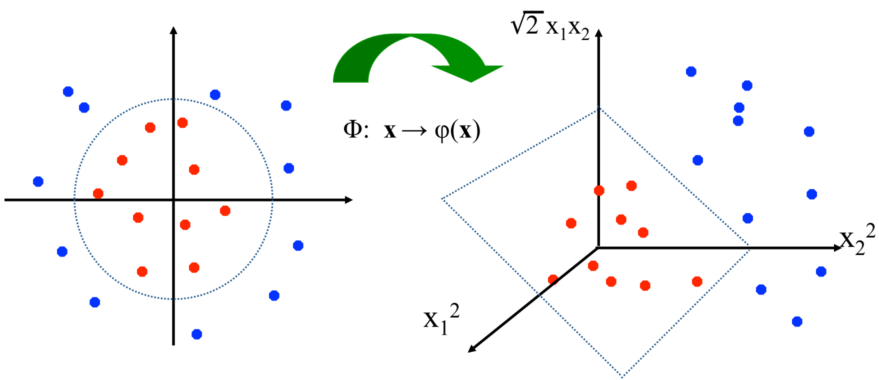


Dr. Nando de Freitas's tutorial slide

Non-linear SVMs: 2D

- The original input space (x) can be mapped to some higher-dimensional feature space (ϕ(**x**))where the training set is separable:

$x=(x_1,x_2)$ $\varphi(\mathbf{x}) =(x_1^2,x_2^2,\sqrt{2}\, x_1x_2)$

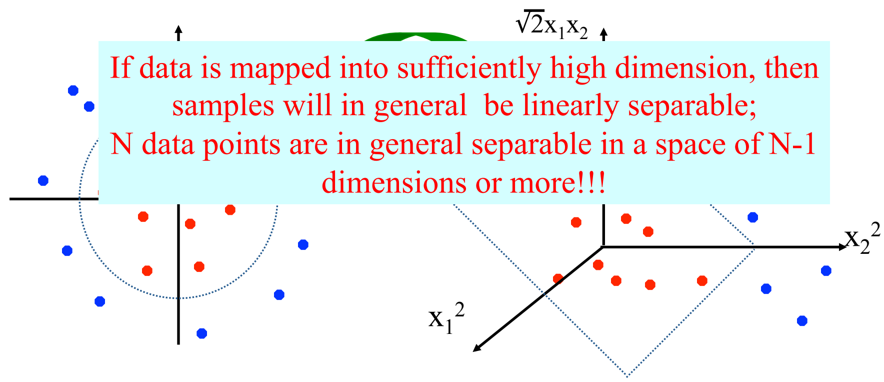


This slide is courtesy of *www.iro.umontreal.ca/~pift6080/documents/papers/**svm_tutorial.ppt***

Non-linear SVMs: 2D

- The original input space (x) can be mapped to some higher-dimensional feature space ($\phi(\mathbf{x})$)where the training set is separable:

$x=(x_1,x_2)$ $\varphi(\mathbf{x}) =(x_1^2,x_2^2,\sqrt{2}x_1x_2)$



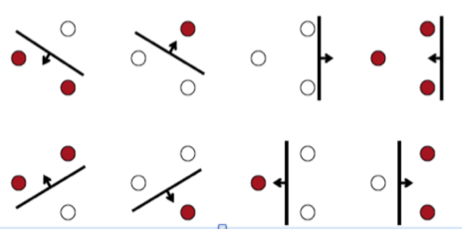
If data is mapped into sufficiently high dimension, then samples will in general be linearly separable;
N data points are in general separable in a space of N-1 dimensions or more!!!

This slide is courtesy of *www.iro.umontreal.ca/~pift6080/documents/papers/**svm**_tutorial.**ppt***

A little bit theory: Vapnik-Chervonenkis (VC) dimension

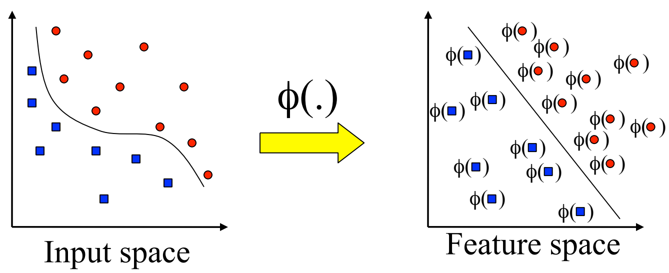
If data is mapped into sufficiently high dimension, then samples will in general be linearly separable;
N data points are in general separable in a space of N-1 dimensions or more!!!

- **VC dimension of the set of oriented lines in R^2 is 3**
  - It can be shown that the VC dimension of the family of oriented separating hyperplanes in R^{N-1} is at least N



Transformation of Inputs

- Possible problems
  - High computation burden due to high-dimensionality
  - Many more parameters
- SVM solves these two issues simultaneously
  - **"Kernel tricks" for efficient computation**
  - **Dual formulation only assigns parameters to samples, not features**



"Kernel tricks" for efficient computation ➔ e.g. Quadratic kernels

- While working in higher dimensions is beneficial, it also increases our running time because of the dot product computation
- However, there is a neat trick we can use
- consider all quadratic terms for $x_1, x_2 \dots x_m$

$$\max_\alpha \sum_i \alpha_i - \sum_{i,j} \alpha_i \alpha_j y_i y_j \Phi(\mathbf{x_i})^T \Phi(\mathbf{x_j})$$
$$\sum_i \alpha_i y_i = 0$$
$$\alpha_i \geq 0 \qquad \forall i$$

m is the number of features in each vector

The √2 term will become clear in the next slide

$$\Phi(x) = \begin{array}{c} 1 \\ \sqrt{2}x_1 \\ \vdots \\ \sqrt{2}x_m \\ x_1^2 \\ \vdots \\ x_m^2 \\ \sqrt{2}x_1x_2 \\ \vdots \\ \sqrt{2}x_{m-1}x_m \end{array}$$

m+1 linear terms

m quadratic terms

m(m-1)/2 pairwise terms

$$K(\mathbf{x}, z) := \Phi(\mathbf{x})^T \Phi(z)$$

Dot product for quadratic kernels

How many operations do we need for the dot product?

$$\Phi(x)^T\Phi(z) = \begin{bmatrix} 1 \\ \sqrt{2}x_1 \\ \vdots \\ \sqrt{2}x_m \\ x_1^2 \\ \vdots \\ x_m^2 \\ \sqrt{2}x_1x_2 \\ \vdots \\ \sqrt{2}x_{m-1}x_m \end{bmatrix} \bullet \begin{bmatrix} 1 \\ \sqrt{2}z_1 \\ \vdots \\ \sqrt{2}z_m \\ z_1^2 \\ \vdots \\ z_m^2 \\ \sqrt{2}z_1z_2 \\ \vdots \\ \sqrt{2}z_{m-1}z_m \end{bmatrix} = \sum_i 2x_iz_i + \sum_i x_i^2z_i^2 + \sum_i \sum_{j=i+1} 2x_ix_jz_iz_j + 1$$

m m m(m-1)/2 **=~ m²**

$K(\mathbf{x}, z) := \Phi(\mathbf{x})^T\Phi(z)$

The kernel trick

How many operations do we need for the dot product?

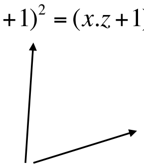
$K(\mathbf{x}, z) := \Phi(\mathbf{x})^T\Phi(z)$

$$\Phi(x)^T\Phi(z) = \sum_i 2x_iz_i + \sum_i x_i^2z_i^2 + \sum_i \sum_{j=i+1} 2x_ix_jz_iz_j + 1$$

m m m(m-1)/2 **=~ m²**

However, we can obtain dramatic savings by noting that

$$\begin{aligned} \Phi(x)^T\Phi(z) = (x^Tz+1)^2 = (x.z+1)^2 &= (x.z)^2 + 2(x.z) + 1 \\ &= (\sum_i x_iz_i)^2 + \sum_i 2x_iz_i + 1 \\ &= \sum_i 2x_iz_i + \sum_i x_i^2z_i^2 + \sum_i \sum_{j=i+1} 2x_ix_jz_iz_j + 1 \end{aligned}$$



We only need m operations!

So, if we define the kernel function as follows, there is no need to carry out $\phi(.)$ explicitly

$$K(\mathbf{x}, z) = (x^Tz+1)^2$$

Where we are

$K(\mathbf{x},z) := \Phi(\mathbf{x})^T \Phi(z)$

Our dual target function:

$$\max_\alpha \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j K(\mathbf{x_i},\mathbf{x_j})$$

$$\sum_i \alpha_i y_i = 0$$

$$\alpha_i \geq 0 \qquad \forall i$$

mn^2 operations at each iteration

To evaluate a new sample x_j we need to compute:

$$w^T \Phi(\mathbf{x_j}) + b = \sum_i \alpha_i y_i K(\mathbf{x_i},\mathbf{x_j}) + b$$

mr operations where *r* are the number of support vectors (α_i>0)

So, if we define the **kernel function** as follows, there is **no need to carry out ϕ(.)** explicitly

$$K(\mathbf{x},z) = (x^T z + 1)^2$$

http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

More examples of kernel functions

- linear: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$.
- polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i{}^T \mathbf{x}_j + r)^d$, $\gamma > 0$.
- radial basis function (RBF): $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$, $\gamma > 0$.
- sigmoid: $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma \mathbf{x}_i{}^T \mathbf{x}_j + r)$.

Here, γ, r, and d are kernel parameters.

Never represent features explicitly
♦ Compute dot products in closed form
Very interesting theory – Reproducing Kernel Hilbert Spaces
☐ Not covered in detail here

$K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ is called the kernel function.

Why do SVMs work?

- If we are using huge features spaces (with kernels) how come we are not overfitting the data?
  - Number of parameters remains the same (and most are set to 0)
  - While we have a lot of input values, at the end we only care about the support vectors and these are usually a small group of samples
  - The minimization (or the maximizing of the margin) function acts as a sort of regularization term leading to reduced overfitting

Today: Review & Practical Guide

- Support Vector Machine (SVM)
  - ✓ Large Margin Linear Classifier
  - ✓ Define Margin (M) in terms of model parameter
  - ✓ Optimization to learn model parameters (w, b)
  - ✓ Non linearly separable case
  - ✓ Optimization with dual form
  - ✓ Nonlinear decision boundary
  - ✓ Practical Guide

Software

- A list of SVM implementation can be found at
  - http://www.kernel-machines.org/software.html

- Some implementation (such as LIBSVM) can handle multi-class classification
- SVMLight is among one of the earliest implementation of SVM
- Several Matlab toolboxes for SVM are also available

Practical Guide to SVM

- From authors of as LIBSVM:
  - A Practical Guide to Support Vector Classification Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, 2003-2010
  - http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

LIBSVM

- http://www.csie.ntu.edu.tw/~cjlin/libsvm/
  - ✓Developed by Chih-Jen Lin etc.
  - ✓Tools for Support Vector classification
  - ✓Also support multi-class classification
  - ✓C++/Java/Python/Matlab/Perl wrappers
  - ✓Linux/UNIX/Windows
  - ✓SMO implementation, fast!!!

A Practical Guide to Support Vector Classification

(a) Data file formats for LIBSVM

- Training.dat

+1 1:0.708333 2:1 3:1 4:-0.320755

-1 1:0.583333 2:-1 4:-0.603774 5:1

+1 1:0.166667 2:1 3:-0.333333 4:-0.433962

-1 1:0.458333 2:1 3:1 4:-0.358491 5:0.374429

…

- Testing.dat

(b) Feature Preprocessing

- (1) Categorical Feature
  - Recommend using m numbers to represent an m-category attribute.
  - Only one of the m numbers is one, and others are zero.
  - For example, a three-category attribute such as {red, green, blue} can be represented as (0,0,1), (0,1,0), and (1,0,0)

A Practical Guide to Support Vector Classification

Feature Preprocessing

- (2) Scaling before applying SVM is very important
  - to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges.
  - to avoid numerical difficulties during the calculation
  - Recommend linearly scaling each attribute to the range [1, +1] or [0, 1].

A Practical Guide to Support Vector Classification

Of course we have to use the same method to scale both training and testing data. For example, suppose that we scaled the first attribute of training data from $[-10, +10]$ to $[-1, +1]$. If the first attribute of testing data lies in the range $[-11, +8]$, we must scale the testing data to $[-1.1, +0.8]$. See Appendix B for some real examples.

If training and testing sets are separately scaled to $[0, 1]$, the resulting accuracy is lower than 70%.

```
$ ../svm-scale -l 0 svmguide4 > svmguide4.scale
$ ../svm-scale -l 0 svmguide4.t > svmguide4.t.scale
$ python easy.py svmguide4.scale svmguide4.t.scale
Accuracy = 69.2308% (216/312) (classification)
```

Using the same scaling factors for training and testing sets, we obtain much better accuracy.

```
$ ../svm-scale -l 0 -s range4 svmguide4 > svmguide4.scale
$ ../svm-scale -r range4 svmguide4.t > svmguide4.t.scale
$ python easy.py svmguide4.scale svmguide4.t.scale
Accuracy = 89.4231% (279/312) (classification)
```

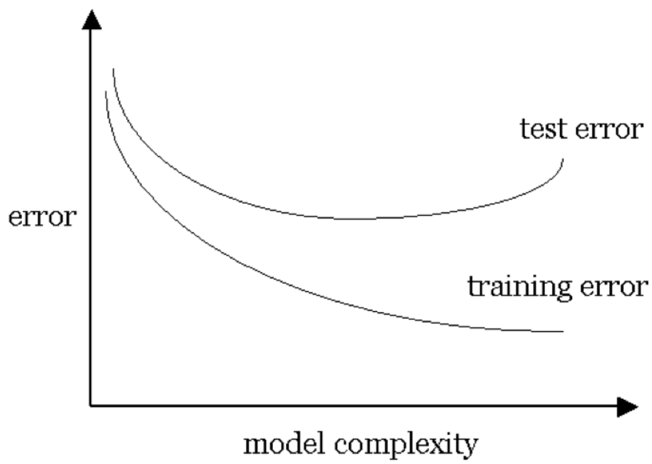

Feature Preprocessing

- (3) missing value
  - Very very tricky !
  - Easy way: to substitute the missing values by the mean value of the variable
  - A little bit harder way: imputation using nearest neighbors
  - Even more complex: e.g. EM based (beyond the scope)

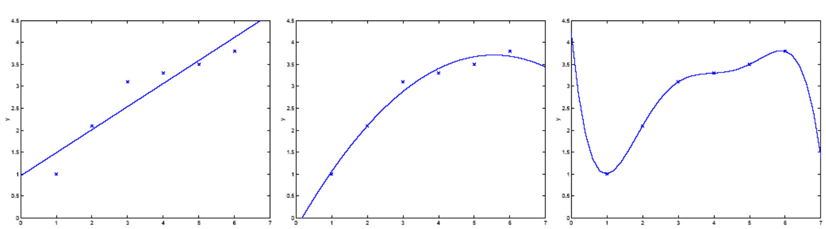
A Practical Guide to Support Vector Classification

(c) Model Selection

Our goal: find the model M which minimizes the test error:



RECAP: Overfitting and underfitting during regression models



$y = \theta_0 + \theta_1 x$ $\qquad$ $y = \theta_0 + \theta_1 x + \theta_2 x^2$ $\qquad$ $y = \sum_{j=0}^{5} \theta_j x^j$

Generalisation: learn function / hypothesis from past data in order to "explain", "predict", "model" or "control" new data examples

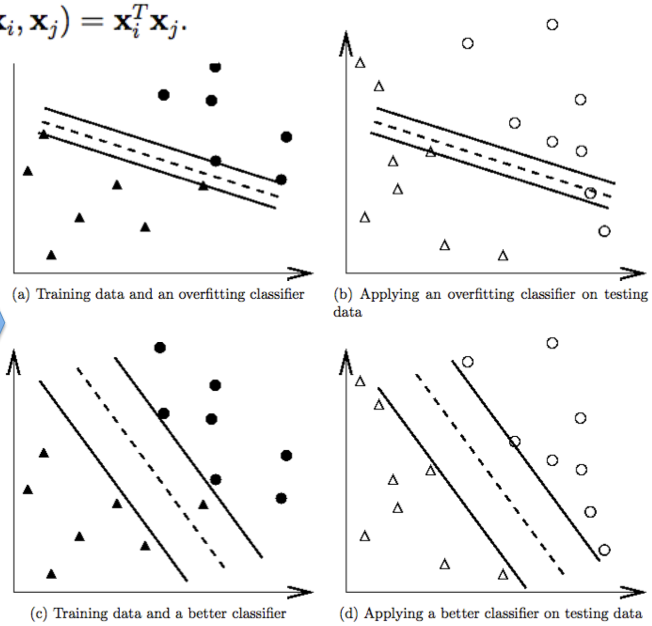
K-fold Cross Validation !!!!

Yanjun Qi / UVA CS 4501-01-6501-07

(c) Model Selection (e.g. for linear kernel)

- linear: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$.

Select the right penalty parameter C



9/26/14 55

Yanjun Qi / UVA CS 4501-01-6501-07

(c) Model Selection

- radial basis function (RBF): $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$, $\gamma > 0$.

  two parameters for an RBF kernel: C and γ

- polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i{}^T \mathbf{x}_j + r)^d$, $\gamma > 0$.

  Three parameters for a polynomial kernel

9/26/14

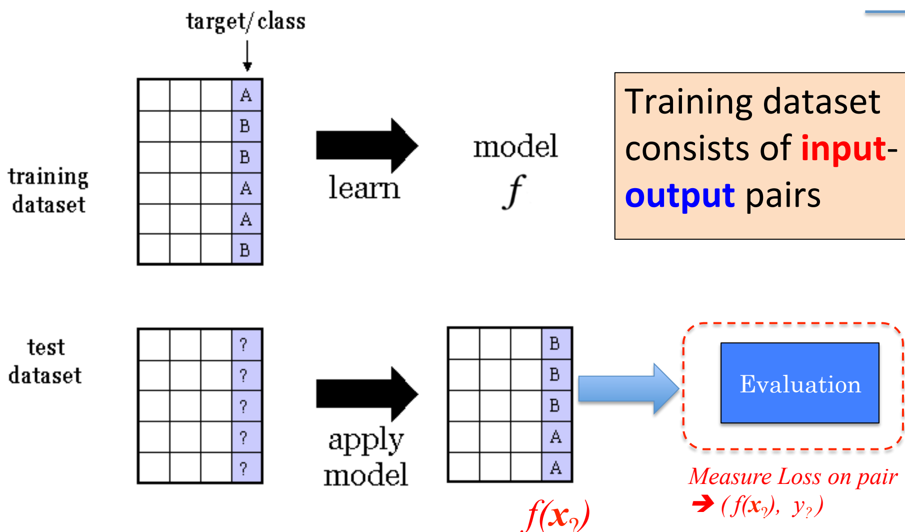
A Practical Guide to Support Vector Classification

56

(d) Pipeline Procedures

- (1) train / test
- (2) k-folds cross validation
- (3) k-CV on train to choose hyperparameter / then test

Evaluation Choice-I: Train and Test



Training dataset consists of **input-output** pairs

Evaluation Choice-II:

Cross Validation

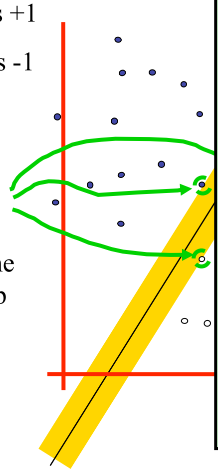
- Problem: don’t have enough data to set aside a test set
- Solution: Each data point is used both as train and test
- Common types:
  - -K-fold cross-validation (e.g. K=5, K=10)
  - -2-fold cross-validation
  - -Leave-one-out cross-validation (LOOCV)

A good practice is : to random shuffle all training sample before splitting

Why Maximum Margin for SVM ?

• denotes +1

° denotes -1



Support Vectors are those datapoints that the margin pushes up against

1. Intuitively this feels safest.
2. If we’ ve made a small error in the location of the boundary (it’ s been jolted in its perpendicular direction) this gives us least chance of causing a misclassification.
3. **LOOCV is easy since the model is immune to removal of any non-support-vector datapoints.**
4. There’ s some theory (using VC dimension) that is related to (but not the same as) the proposition that this is a good thing.
5. Empirically it works very very well.

Yanjun Qi / UVA CS 4501-01-6501-07

Many beginners use the following procedure now:

- Transform data to the format of an SVM package
- Randomly try a few kernels and parameters
- Test

Basic solution For HW2-Q2

We propose that beginners try the following procedure first:

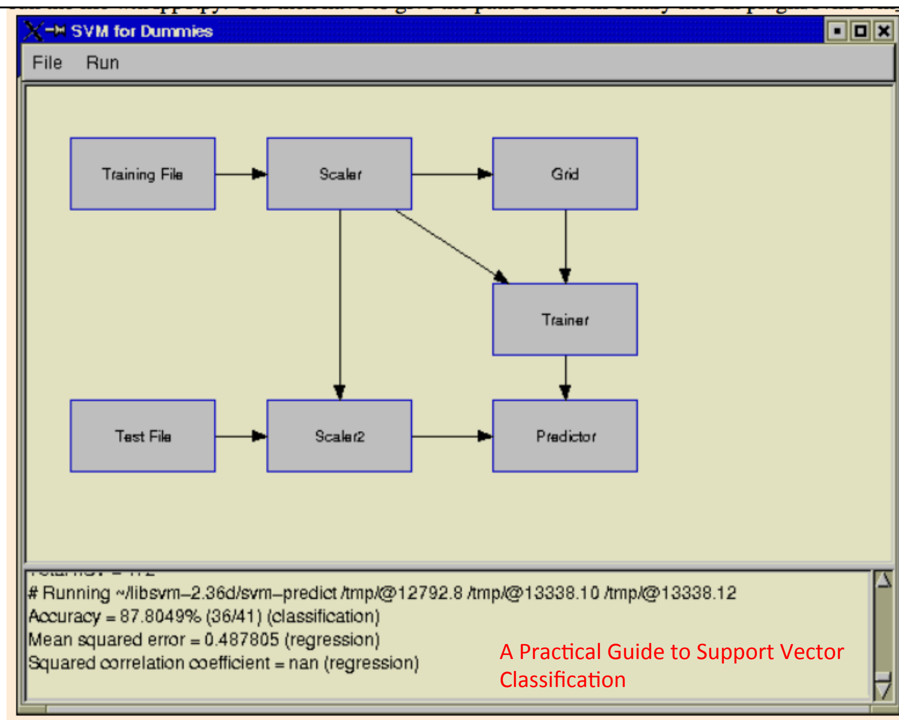
- Transform data to the format of an SVM package
- Conduct simple scaling on the data
- Consider the RBF kernel $K(\mathbf{x}, \mathbf{y}) = e^{-\gamma\|\mathbf{x}-\mathbf{y}\|^2}$
- Use cross-validation to find the best parameter C and γ
- Use the best parameter C and γ to train the whole training set[5]
- Test

more advanced solution For HW2-Q2

Evaluation Choice-III

A Practical Guide to Support Vector Classification

9/26/14 61



A Practical Guide to Support Vector Classification

Today: Review & Practical Guide

- Support Vector Machine (SVM)
  - ✓ Large Margin Linear Classifier
  - ✓ Define Margin (M) in terms of model parameter
  - ✓ Optimization to learn model parameters (w, b)
  - ✓ Non linearly separable case
  - ✓ Optimization with dual form
  - ✓ Nonlinear decision boundary
  - ✓ Practical Guide
    - ✓ File format / LIBSVM
    - ✓ Feature preprocsssing
    - ✓ Model selection
    - ✓ Pipeline procedure

References

- Big thanks to Prof. Ziv Bar-Joseph @ CMU for allowing me to reuse some of his slides
- Prof. Andrew Moore @ CMU's slides
- Elements of Statistical Learning, by Hastie, Tibshirani and Friedman