UVA CS 6316 – Fall 2015 Graduate: Machine Learning

Lecture 8: Supervised Classification with Support Vector Machine

Dr. Yanjun Qi

University of Virginia

Department of Computer Science

Where we are ? ➔ Five major sections of this course

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

Today

- ❑ Supervised Classification
- ❑ Support Vector Machine (SVM)

e.g. SUPERVISED LEARNING

- Find function to map input space X to output space Y $f : X \longrightarrow Y$
- So that the difference between y and $f(x)$ of each example $\boldsymbol{x}$ is small.

e.g.

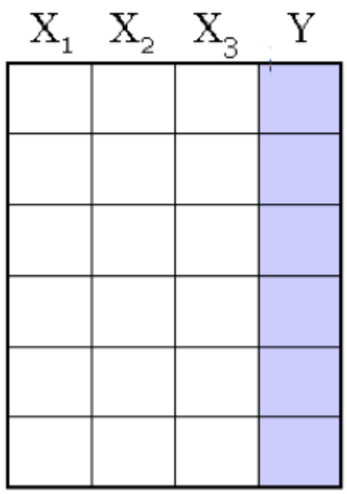
x	I believe that this book is not at all helpful since it does not explain thoroughly the material . it just provides the reader with tables and calculations that sometimes are not easily understood …

Input X : e.g. a piece of English text

→

y	-1

Output Y: {1 / Yes , -1 / No }
e.g. Is this a positive product review ?



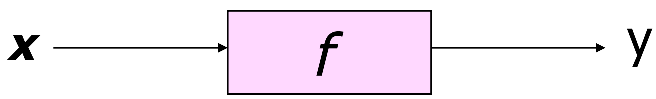
A Dataset for classification

$$f : X \longrightarrow Y$$

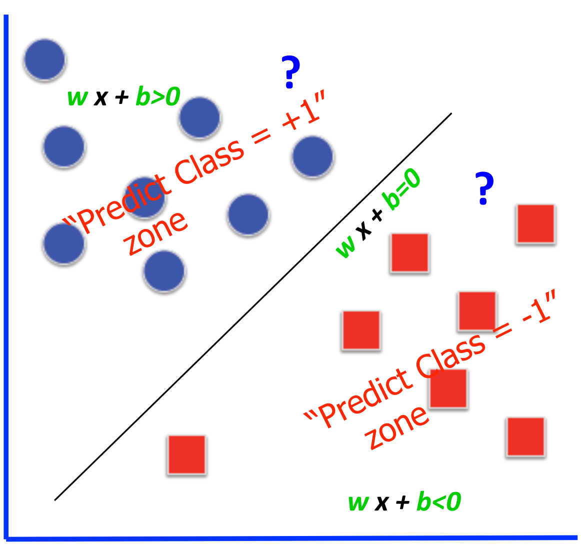
Output Class: categorical variable

- **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]

e.g. SUPERVISED Linear Binary Classifier



$$f(\mathbf{x}, \mathbf{w}, b) = sign(\mathbf{w}\ \mathbf{x} + b)$$



denotes +1 point

denotes -1 point

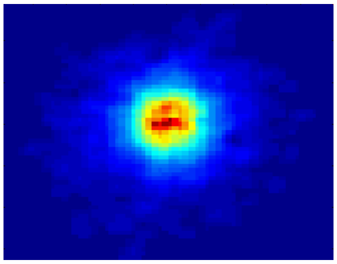
? denotes future points

Courtesy slide from Prof. Andrew Moore's tutorial

Application 1: Classifying Galaxies

Courtesy: http://aps.umn.edu

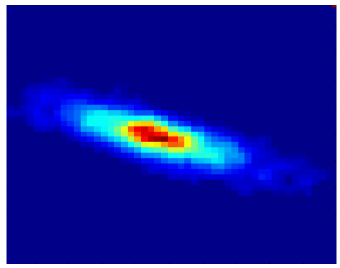
Early



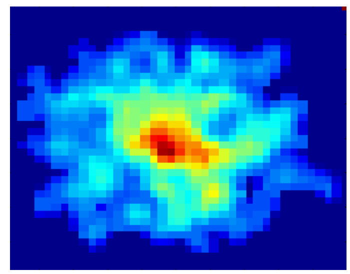
Class:
- Stages of Formation

Attributes:
- Image features,
- Characteristics of light waves received, etc.

Intermediate



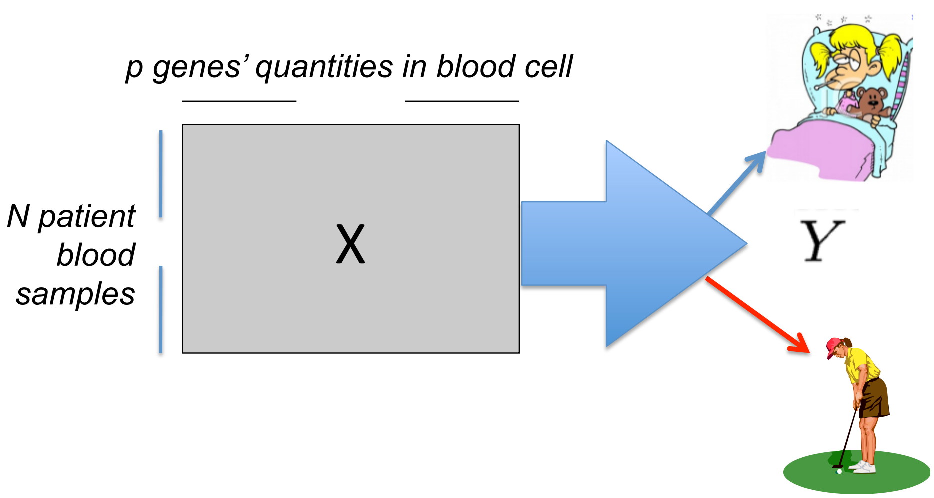
Late



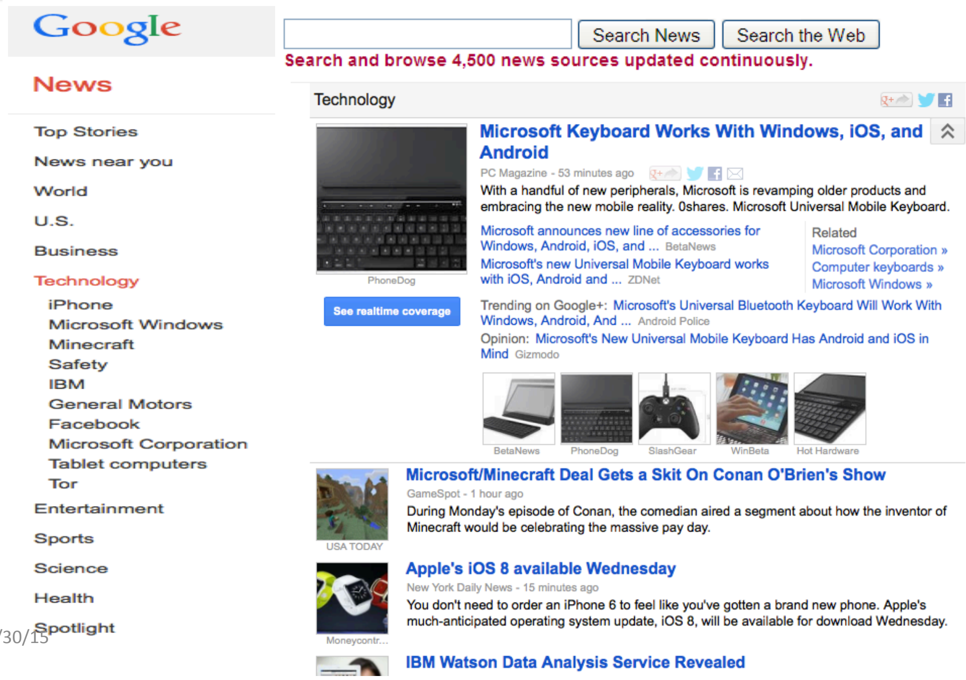
Data Size:
- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

From [Berry & Linoff] Data Mining Techniques, 1997

Application 2: **Cancer Classification using gene expression**



Application 3: –
Text Documents, e.g. *Google News*



Text Document Representation

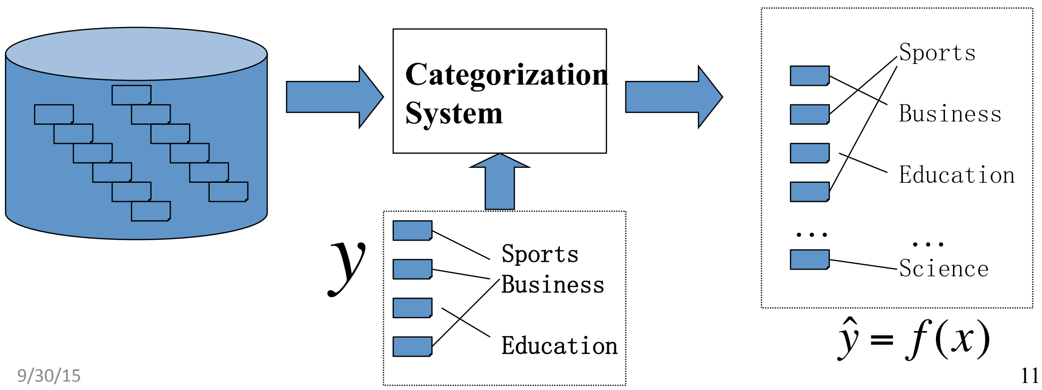
- Each document becomes a `term' vector,
  - each term is an (attribute) of the vector,
  - the value of each describes the number of times the corresponding term occurs in the document.

Bag of 'words'

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Text Categorization

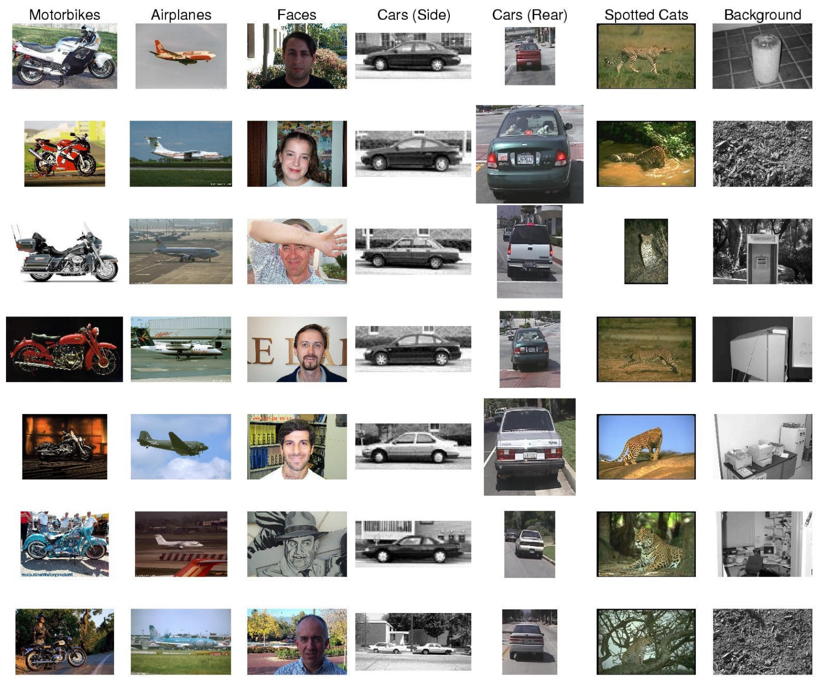
- Pre-given categories and labeled document examples (Categories may form hierarchy)
- Classify new documents
- A standard supervised learning problem



Examples of Text Categorization

- News article classification
- Meta-data annotation
- Automatic Email sorting
- Web page classification

Application 4: – Objective recognition / Image Labeling (Label Images into predefined classes)

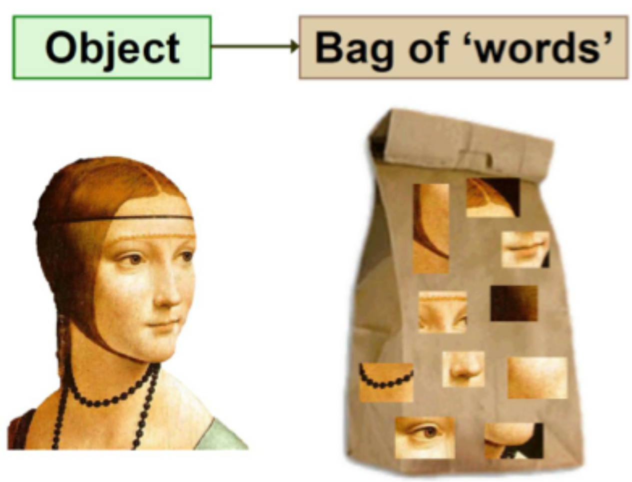


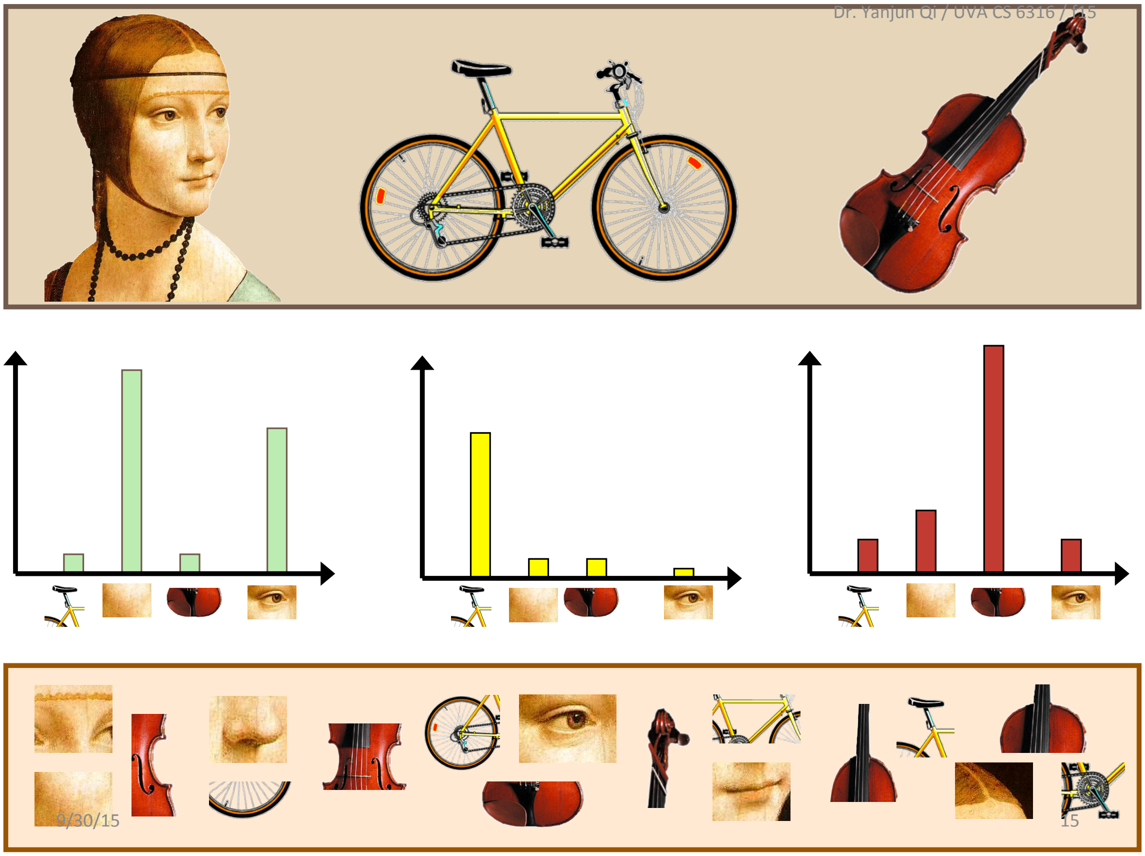
Dr. Yanjun Qi / UVA CS 6316 / f15

Image Representation for – Objective recognition

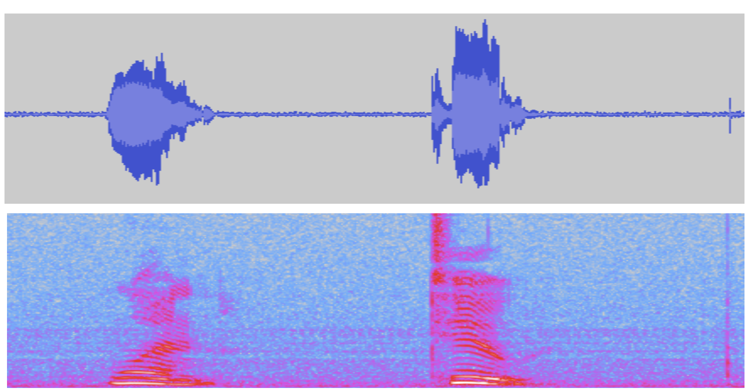
- Image representation ➔ bag of "visual words"

- An object image: histogram of visual vocabulary – a numerical vector of D dimensions.





Application 5: – Audio Classification



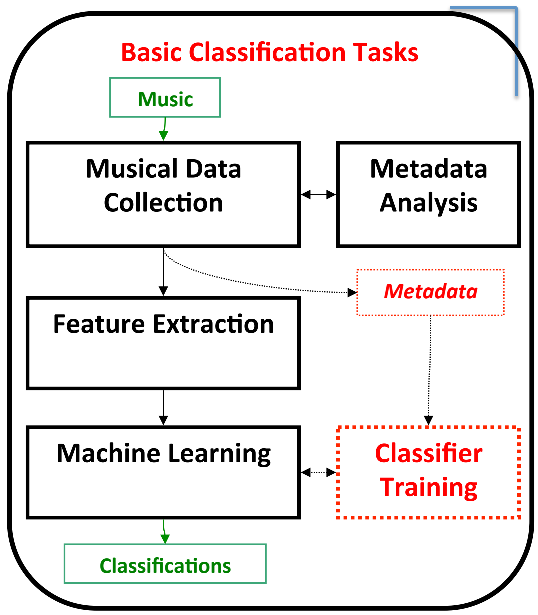
- Real-life applications:
  - Customer service phone routing
  - Voice recognition software

Music Information Retrieval Systems e.g., Automatic Music Classification

- Many areas of research in music information retrieval (MIR) involve using computers to classify music in various ways
  - Genre or style classification
  - Mood classification
  - Performer or composer identification
  - Music recommendation
  - Playlist generation
  - Hit prediction
  - Audio to symbolic transcription
  - etc.
- Such areas often share similar central procedures

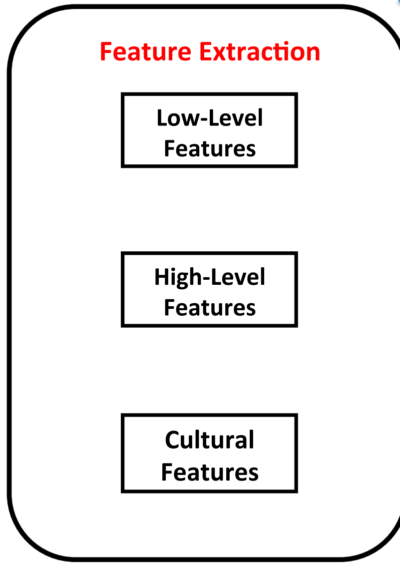
Music Information Retrieval Systems e.g., Automatic Music Classification

- Musical data collection
  - The instances (basic entities) to classify
  - Audio recordings, scores, cultural data, etc.
- Feature extraction
  - Features represent characteristic information about instances
  - Must provide sufficient information to segment instances among classes (categories)
- Machine learning
  - Algorithms ("classifiers" or "learners") learn to associate feature patterns of instances with their classes



Audio, Types of features

- Low-level
  - Associated with signal processing and basic auditory perception
  - e.g. spectral flux or RMS
  - Usually not intuitively musical
- High-level
  - Musical abstractions
  - e.g. meter or pitch class distributions
- Cultural
  - Sociocultural information outside the scope of auditory or musical content
  - e.g. playlist co-occurrence or purchase correlations



Where we are ? ➔ 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., Bayesian networks

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

ICML '06 Proceedings of the 23rd international conference on Machine learning

A study comparing Classifiers

An Empirical Comparison of Supervised Learning Algorithms

Rich Caruana CARUANA@CS.CORNELL.EDU
Alexandru Niculescu-Mizil ALEXN@CS.CORNELL.EDU
Department of Computer Science, Cornell University, Ithaca, NY 14853 USA

Abstract

A number of supervised learning methods have been introduced in the last decade. Unfortunately, the last comprehensive empirical evaluation of supervised learning was the Statlog Project in the early 90's. We present a large-scale empirical comparison between ten supervised learning methods: SVMs, neural nets, logistic regression, naive bayes, memory-based learning, random forests, decision trees, bagged trees, boosted trees, and boosted stumps. We also examine the effect that calibrating the models via Platt Scaling and Isotonic Regression has on their performance. An important aspect of our study is

This paper presents results of a large-scale empirical comparison of ten supervised learning algorithms using eight performance criteria. We evaluate the performance of SVMs, neural nets, logistic regression, naive bayes, memory-based learning, random forests, decision trees, bagged trees, boosted trees, and boosted stumps on eleven binary classification problems using a variety of performance metrics: accuracy, F-score, Lift, ROC Area, average precision, precision/recall break-even point, squared error, and cross-entropy. For each algorithm we examine common variations, and thoroughly explore the space of parameters. For example, we compare ten decision tree styles, neural nets of many sizes, SVMs with many kernels, etc.

Because some of the performance metrics we examine

Dr. Yanjun Qi / UVA CS 6316 / f15

A study comparing Classifiers → 11 binary classification problems

PROBLEM	#ATTR	TRAIN SIZE	TEST SIZE	%POZ
ADULT	14/104	5000	35222	25%
BACT	11/170	5000	34262	69%
COD	15/60	5000	14000	50%
CALHOUS	9	5000	14640	52%
COV_TYPE	54	5000	25000	36%
HS	200	5000	4366	24%
LETTER.P1	16	5000	14000	3%
LETTER.P2	16	5000	14000	53%
MEDIS	63	5000	8199	11%
MG	124	5000	12807	17%
SLAC	59	5000	25000	50%

A study comparing Classifiers

→ 11 binary classification problems / 8 metrics

Table 2. Normalized scores for each learning algorithm by metric (average over eleven problems)

MODEL	CAL	ACC	FSC	LFT	ROC	APR	BEP	RMS	MXE	MEAN	OPT-SEL
BST-DT	PLT	.843*	.779	**.939**	**.963**	**.938**	.929*	**.880**	**.896**	**.896**	**.917**
RF	PLT	.872*	.805	.934*	.957	.931	**.930**	.851	.858	.892	.898
BAG-DT	–	.846	.781	.938*	.962*	.937*	.918	.845	.872	.887*	.899
BST-DT	ISO	.826*	.860*	.929*	.952	.921	.925*	.854	.815	.885	.917*
RF	–	**.872**	.790	.934*	.957	.931	**.930**	.829	.830	.884	.890
BAG-DT	PLT	.841	.774	.938*	.962*	.937*	.918	.836	.852	.882	.895
RF	ISO	.861*	**.861**	.923	.946	.910	.925	.836	.776	.880	.895
BAG-DT	ISO	.826	.843*	.933*	.954	.921	.915	.832	.791	.877	.894
SVM	PLT	.824	.760	.895	.938	.898	.913	.831	.836	.862	.880
ANN	–	.803	.762	.910	.936	.892	.899	.811	.821	.854	.885
SVM	ISO	.813	.836*	.892	.925	.882	.911	.814	.744	.852	.882
ANN	PLT	.815	.748	.910	.936	.892	.899	.783	.785	.846	.875
ANN	ISO	.803	.836	.908	.924	.876	.891	.777	.718	.842	.884
BST-DT	–	.834*	.816	**.939**	**.963**	**.938**	.929*	.598	.605	.828	.851
KNN	PLT	.757	.707	.889	.918	.872	.872	.742	.764	.815	.837
KNN	–	.756	.728	.889	.918	.872	.872	.729	.718	.810	.830
KNN	ISO	.755	.758	.882	.907	.854	.869	.738	.706	.809	.844
BST-STMP	PLT	.724	.651	.876	.908	.853	.845	.716	.754	.791	.808
SVM	–	.817	.804	.895	.938	.899	.913	.514	.467	.781	.810
BST-STMP	ISO	.709	.744	.873	.899	.835	.840	.695	.646	.780	.810
BST-STMP	–	.741	.684	.876	.908	.853	.845	.394	.382	.710	.726
DT	ISO	.648	.654	.818	.838	.756	.778	.590	.589	.709	.774

Today

- ❑ Supervised Classification
- ❑ Support Vector Machine (SVM)
  - ✓ History of 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
  - ✓ Multiclass SVM

History of SVM

Young / theoretically sound / Impactful

- SVM is inspired from statistical learning theory [3]
- SVM was first introduced in 1992 [1]
- SVM becomes popular because of its success in handwritten digit recognition (1994)
  - 1.1% test error rate for SVM. This is the same as the error rates of a carefully constructed neural network, LeNet 4.
    - Section 5.11 in [2] or the discussion in [3] for details
- SVM is now regarded as an important example of "kernel methods", arguably the hottest area in machine learning 15 years ago

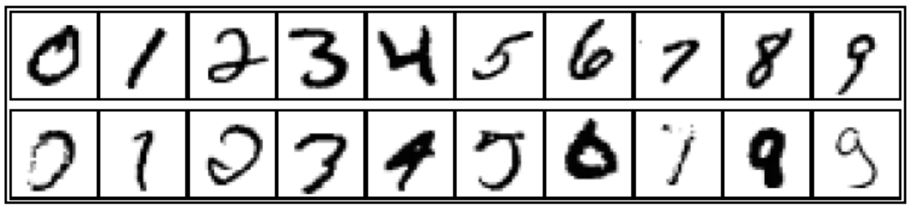
[1] B.E. Boser *et al.* A Training Algorithm for Optimal Margin Classifiers. Proceedings of the Fifth Annual Workshop on Computational Learning Theory 5 144-152, Pittsburgh, 1992.
[2] L. Bottou *et al.* Comparison of classifier methods: a case study in handwritten digit recognition. Proceedings of the 12th IAPR International Conference on Pattern Recognition, vol. 2, pp. 77-82, 1994.
[3] V. Vapnik. The Nature of Statistical Learning Theory. 2nd edition, Springer, 1999.

Applications of SVMs

- Computer Vision
- Text Categorization
- Ranking (e.g., Google searches)
- Handwritten Character Recognition
- Time series analysis
- Bioinformatics
- ……….

→Lots of very successful applications!!!

Handwritten digit recognition

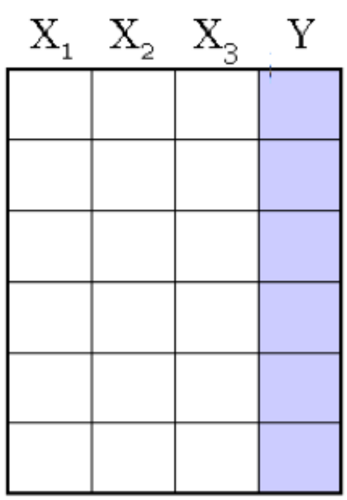


3-nearest-neighbor = 2.4% error
400–300–10 unit MLP = 1.6% error
LeNet: 768–192–30–10 unit MLP = 0.9% error

In 90s, SVM achieves the

best (kernel machines, vision algorithms) ≈ 0.6% error

A Dataset for binary classification



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

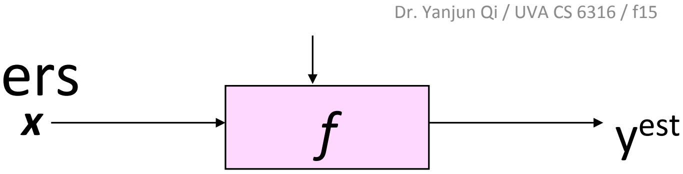
Output as Binary Class: only two possibilities

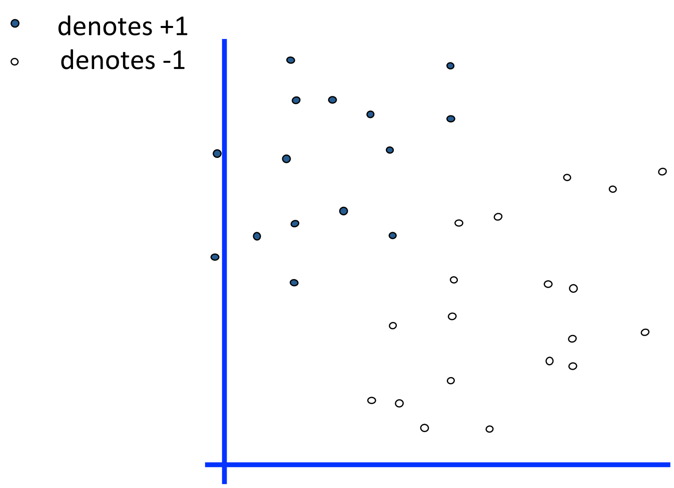
- **Data**/*points/instances/examples/samples/records*: [rows]
- **Features**/*attributes/dimensions/independent variables/covariates/predictors/regressors*: [columns, except the last]
- **Target**/*outcome/response/label/dependent variable*: special column to be predicted [last column]

Today

- ❑ Supervised Classification
- ❑ Support Vector Machine (SVM)
  - ✓ History of 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
  - ✓ Multiclass SVM

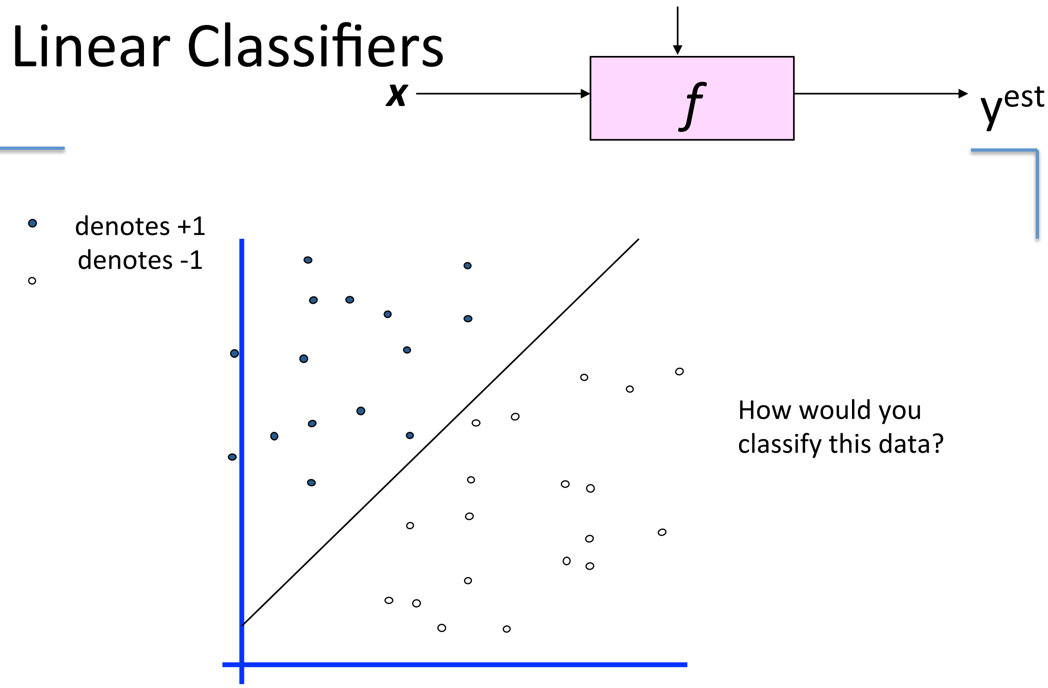
Linear Classifiers





How would you classify this data?

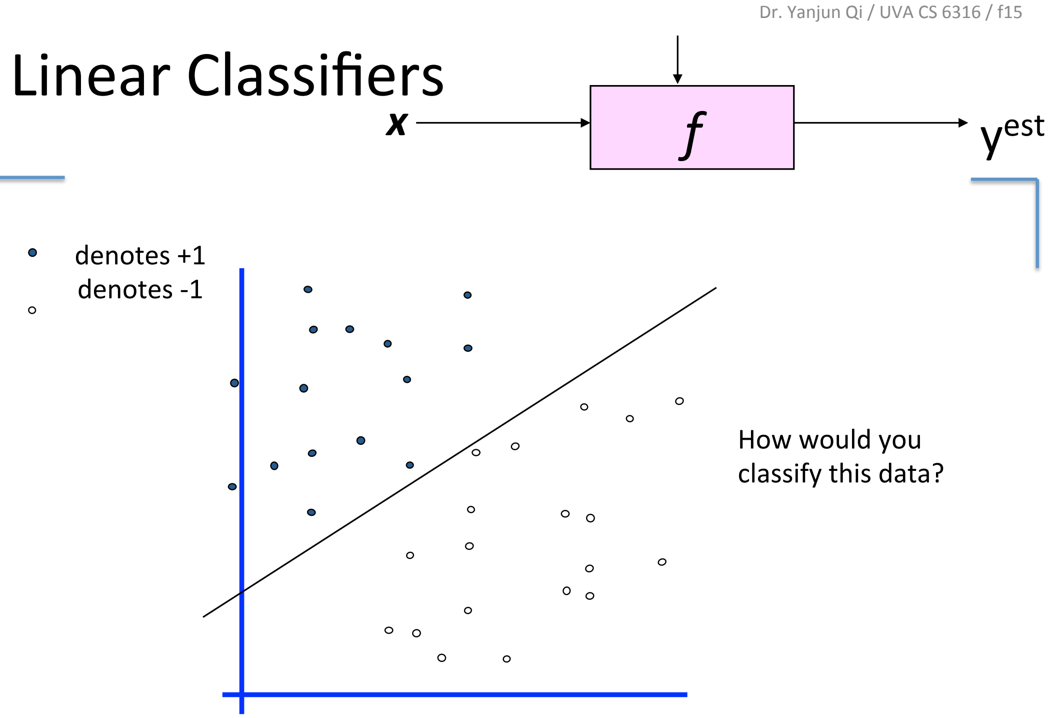
Linear Classifiers



How would you classify this data?

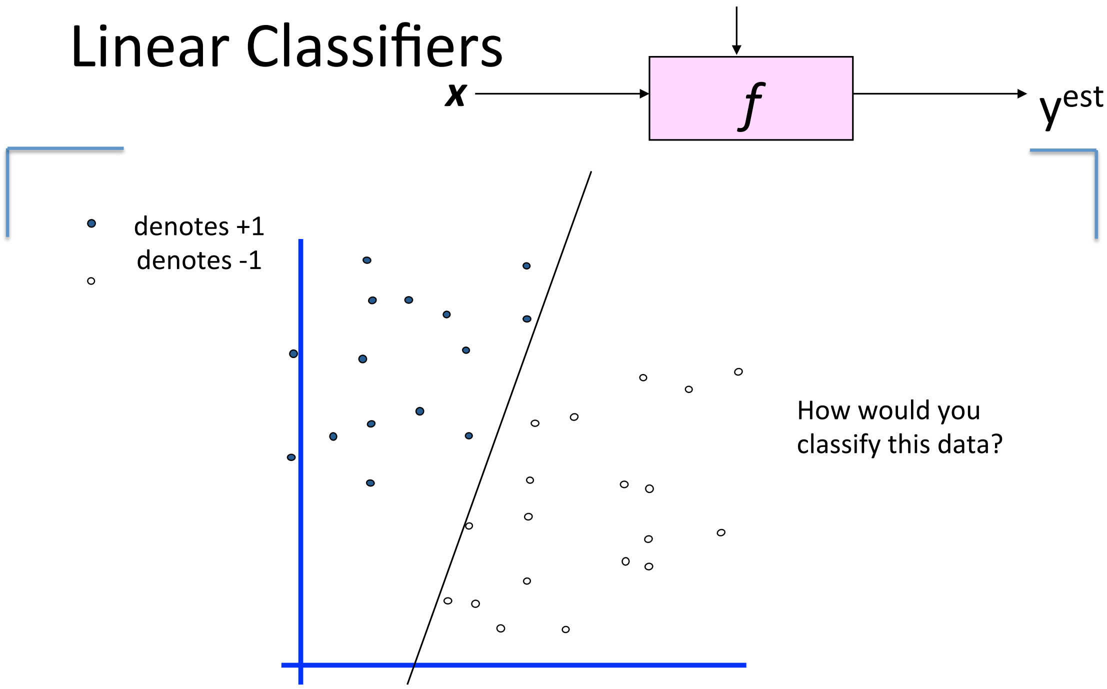
Dr. Yanjun Qi / UVA CS 6316 / f15

Linear Classifiers



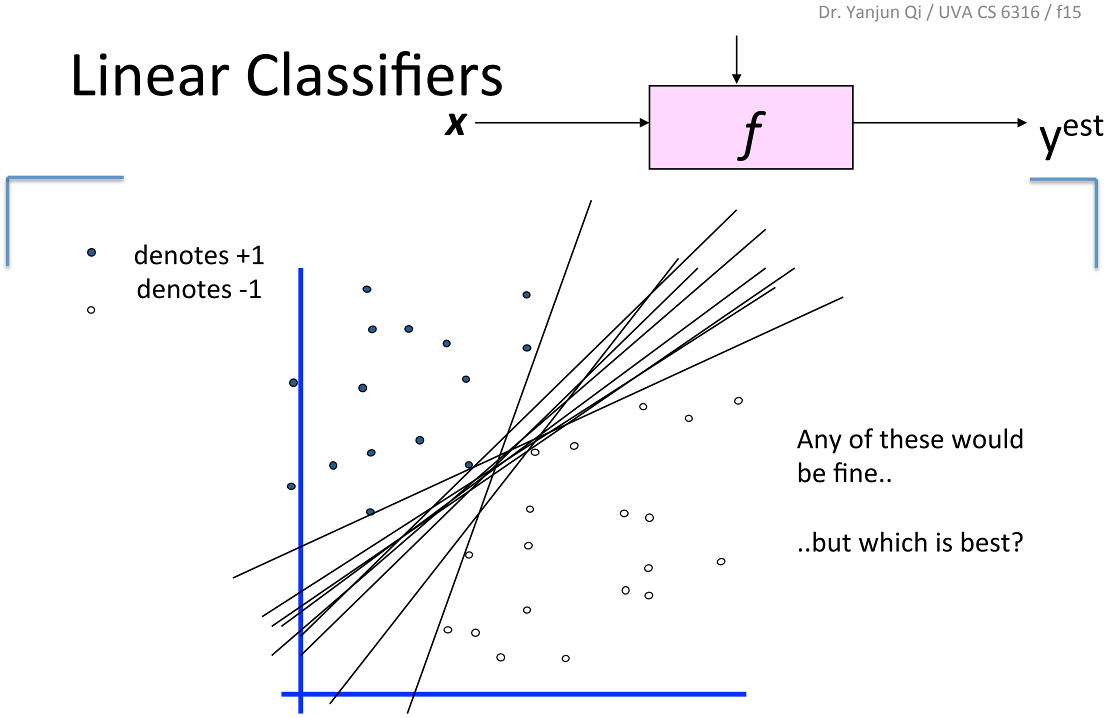
How would you classify this data?

Linear Classifiers



How would you classify this data?

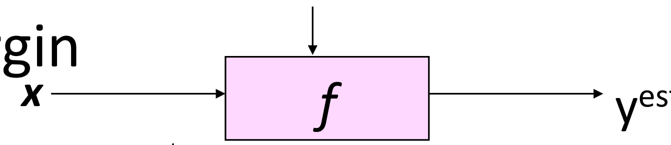
Linear Classifiers

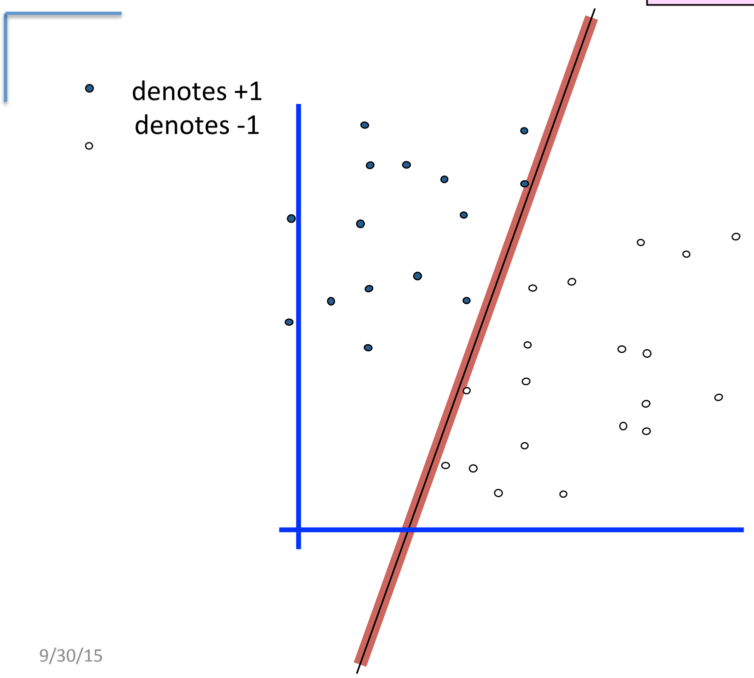


Any of these would be fine..

..but which is best?

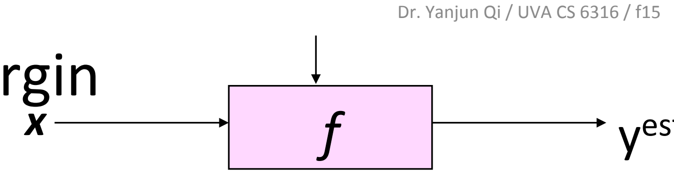
Classifier Margin

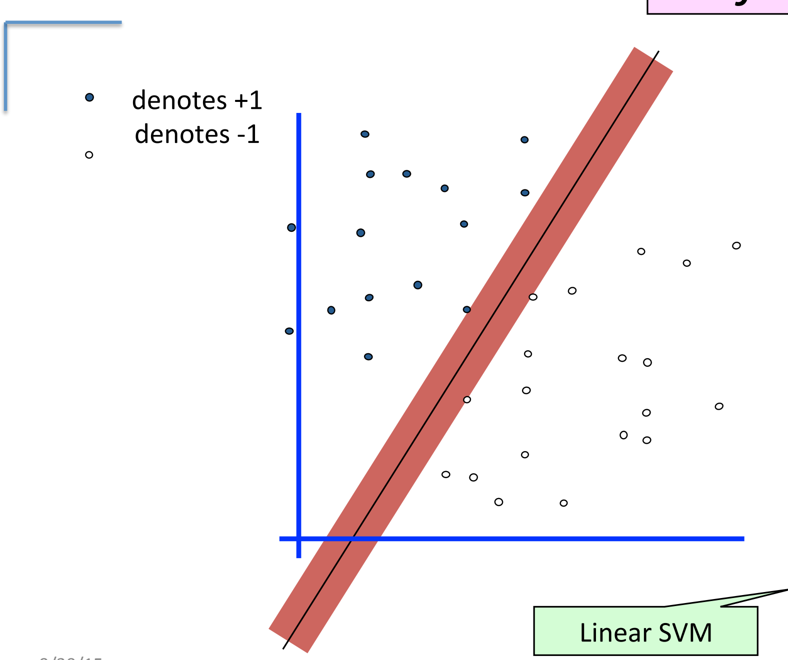




Define the margin of a linear classifier as the width that the boundary could be increased by before hitting a datapoint.

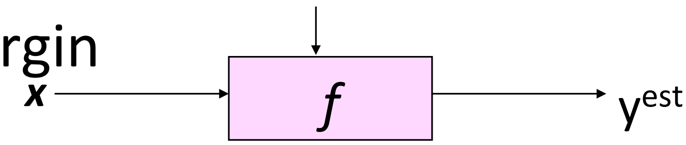
Maximum Margin

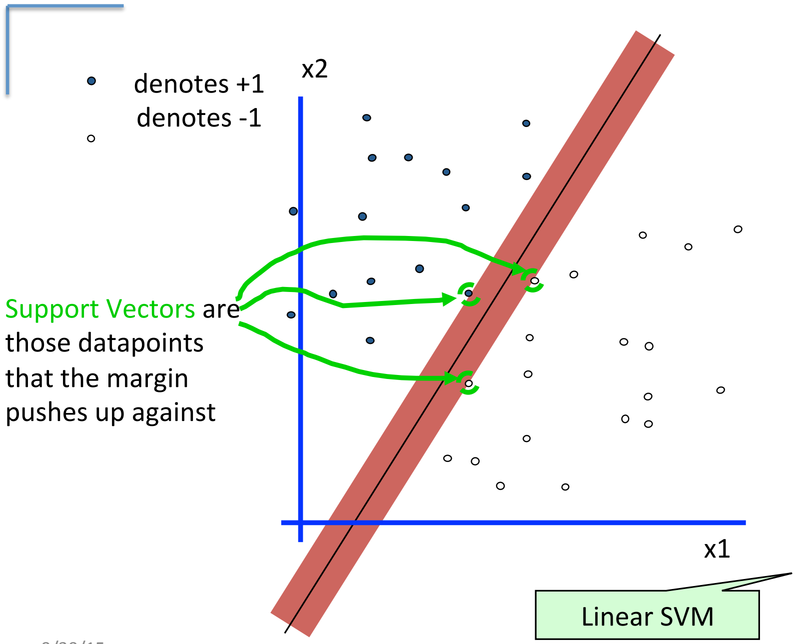




The maximum margin linear classifier is the linear classifier with the, maximum margin.
This is the simplest kind of SVM (Called an LSVM)

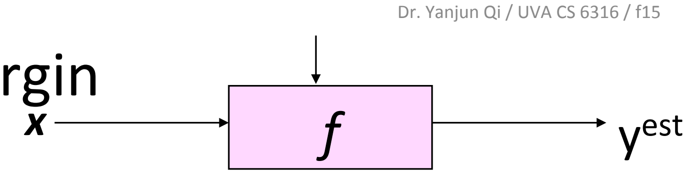
Maximum Margin



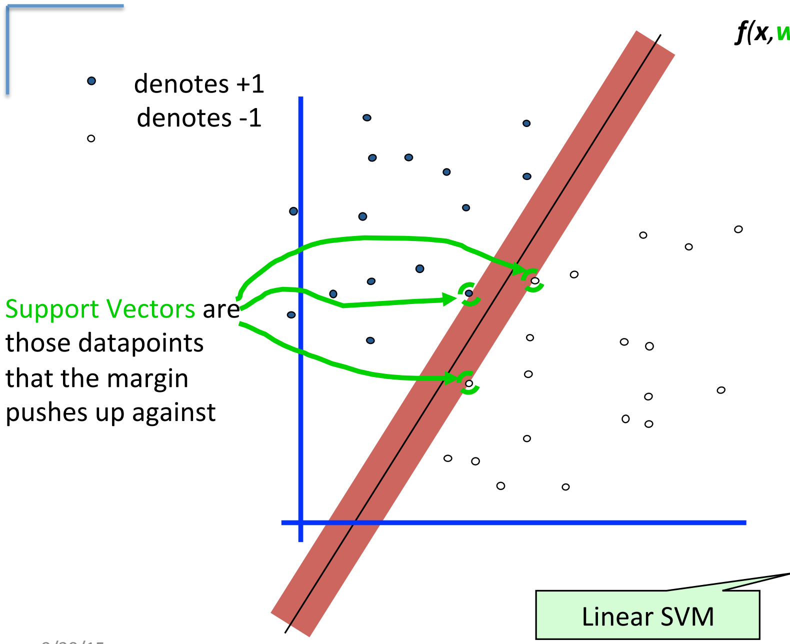


The maximum margin linear classifier is the linear classifier with the, maximum margin.
This is the simplest kind of SVM (Called an LSVM)

Maximum Margin



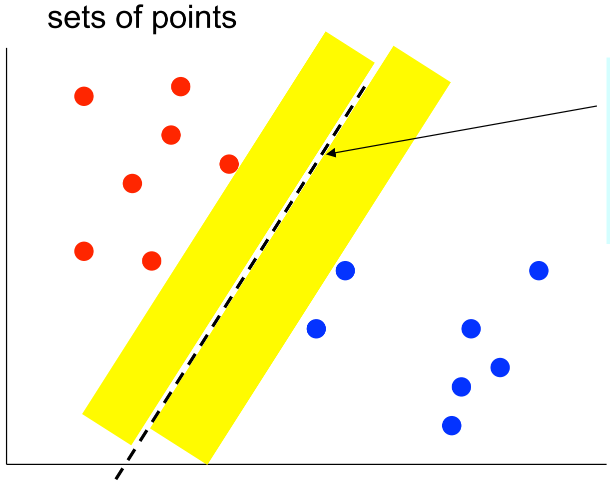
$f(\mathbf{x},\mathbf{w},b) = sign(\mathbf{w}^T\mathbf{x} + b)$



The maximum margin linear classifier is the linear classifier with the, maximum margin.
This is the simplest kind of SVM (Called an LSVM)

Max margin classifiers

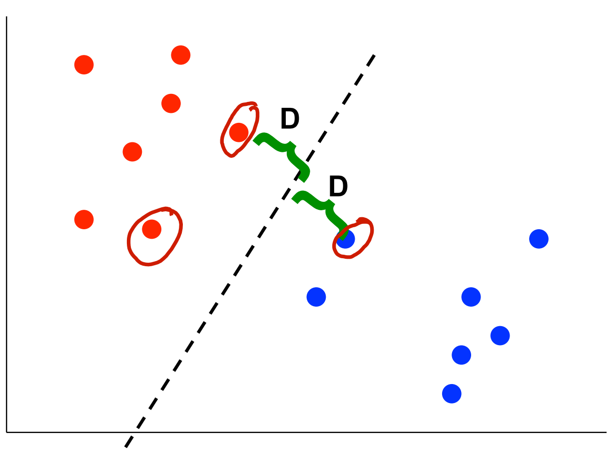
- Instead of fitting all points, focus on boundary points
- Learn a boundary that leads to the largest margin from both sets of points



From all the possible boundary lines, this leads to the largest margin on both sides

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

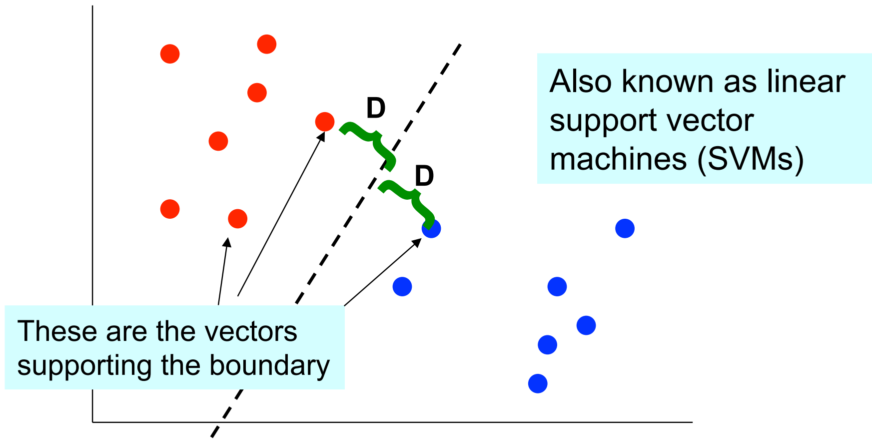


Why?

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

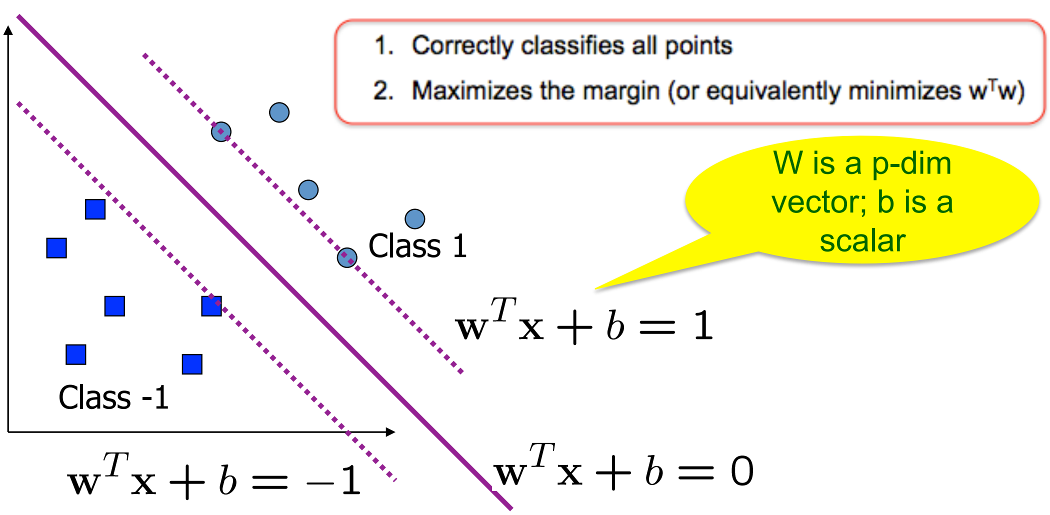
Max margin classifiers

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

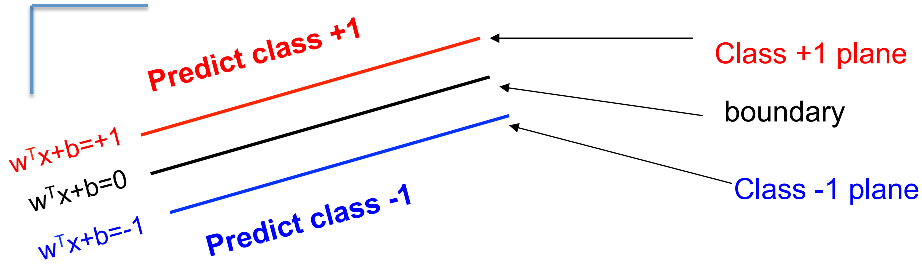


Max-margin & Decision Boundary

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

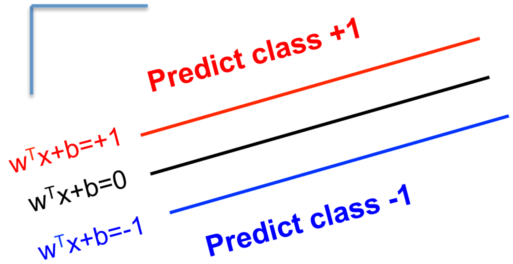


Specifying a max margin classifier



Classify as +1	if	$w^Tx+b >= 1$
Classify as -1	if	$w^Tx+b <= -1$
Undefined	if	$-1 < w^Tx+b < 1$

Specifying a max margin classifier



Is the linear separation assumption realistic?

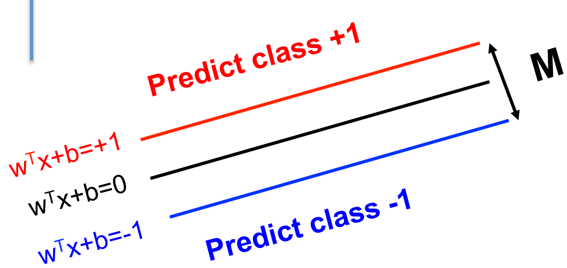
We will deal with this shortly, but lets assume it for now

Classify as +1	if	$w^Tx+b >= 1$
Classify as -1	if	$w^Tx+b <= -1$
Undefined	if	$-1 < w^Tx+b < 1$

Today

- ❑ Supervised Classification
- ❑ Support Vector Machine (SVM)
  - ✓ History of 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
  - ✓ Multiclass SVM

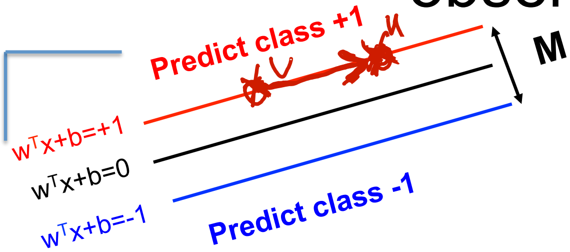
Maximizing the margin



Classify as +1 if $w^Tx+b >= 1$
Classify as -1 if $w^Tx+b <= -1$
Undefined if $-1 < w^Tx+b < 1$

- Lets define the width of the margin by M
- How can we encode our goal of maximizing M in terms of our parameters (w and b)?
- Lets start with a few obsevrations

Maximizing the margin: observation-1



Classify as +1 if $w^Tx+b >= 1$
Classify as -1 if $w^Tx+b <= -1$
Undefined if $-1 < w^Tx+b < 1$

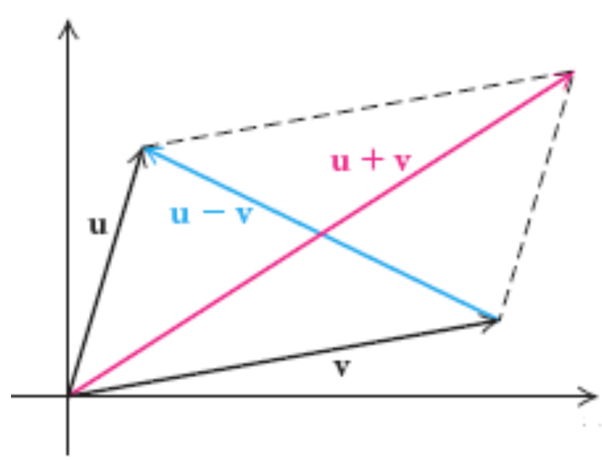
- **Observation 1: the vector w is orthogonal to the +1 plane**
- Why?

$$W^T(u-v) = W^Tu - W^Tv = 0$$

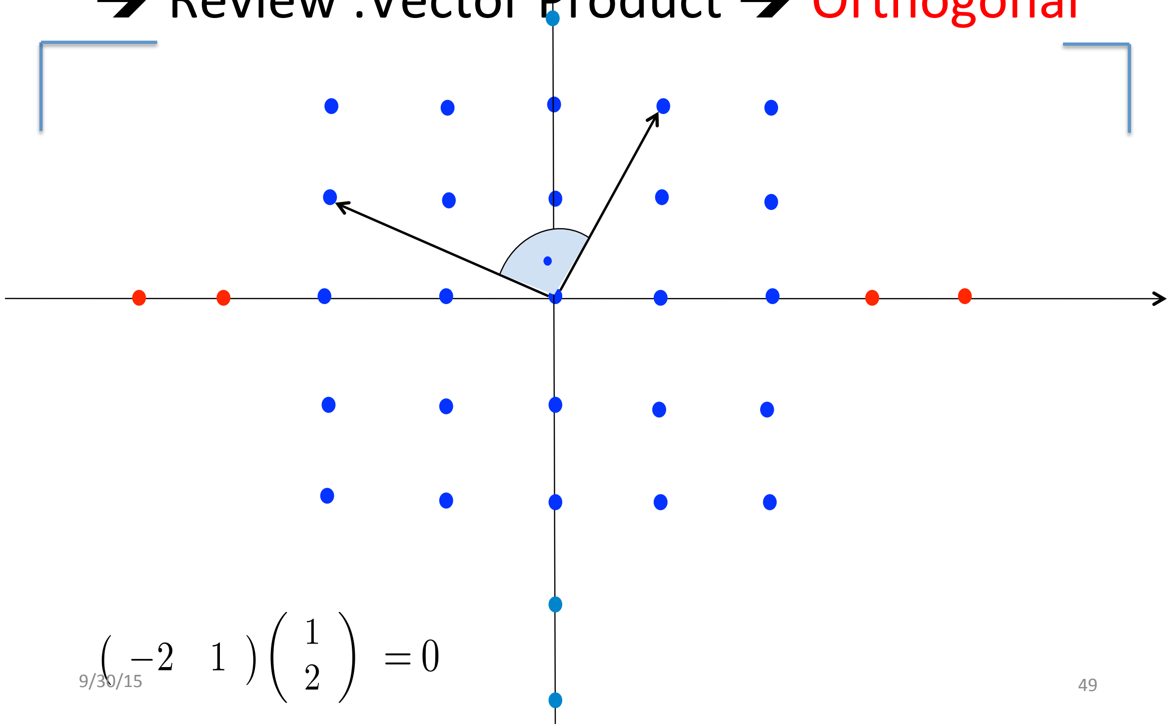
Let u and v be two points on the +1 plane, then for the vector defined by u and v we have $w^T(u-v) = 0$

Corollary: the vector w is orthogonal to the -1 plane

Observation 1
➔ Review :Vector Subtraction



Observation 1 → Review :Vector Product → Orthogonal



$$\begin{pmatrix} -2 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = 0$$

Observation 1 → Review : Vector Product, Orthogonal, and Norm

For two vectors *x* and *y*,

$$x^T y$$

is called the *(inner) vector product.*

x and *y* are called *orthogonal* if

$$x^T y = 0$$

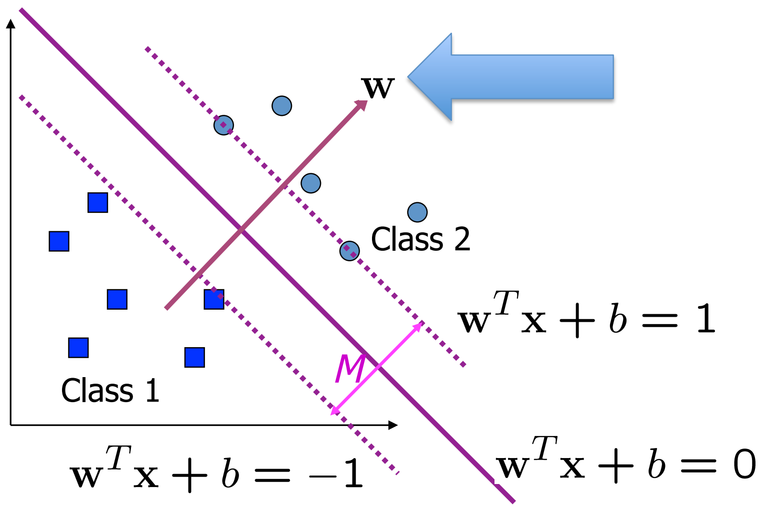
The square root of the product of a vector with itself,

$$\sqrt{x^T x}$$

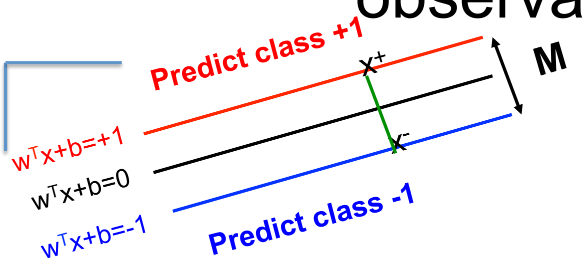
is called the *2-norm* ($|x|_2$), can also write as $|x|$

Maximizing the margin: observation-1

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



Maximizing the margin: observation-2



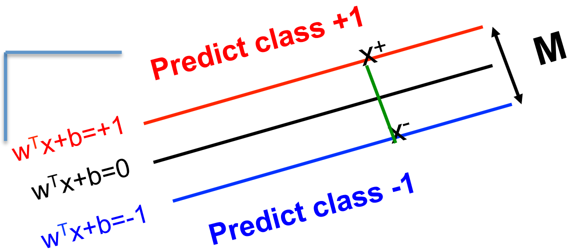
Classify as +1	if	$w^Tx+b >= 1$
Classify as -1	if	$w^Tx+b <= -1$
Undefined	if	$-1 < w^Tx+b < 1$

- 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



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

We can now define M in terms of w and b

$$M = |x^+ - x^-| = |\lambda w| = \lambda |w| = \lambda \sqrt{w^T w} = \frac{2}{w^T w}\sqrt{w^T w}$$

$$\left[= \frac{2}{\sqrt{w^T w}} \right]$$

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

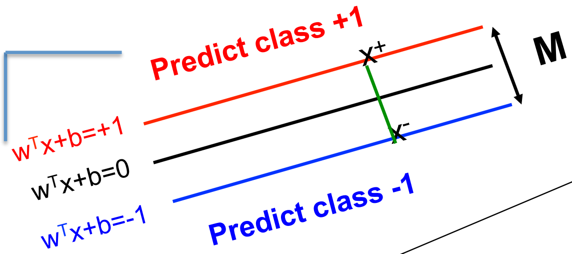
$$w^T(\lambda w + x^-) + b = +1$$

$$\lambda w^T w + \underbrace{w^T x^- + b} = 1$$

$$\lambda w^T w = 2$$

$$\Rightarrow \lambda = \frac{2}{w^T w}$$

Putting it together



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

We can now define M in terms of w and b

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

=>

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

=>

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

=>

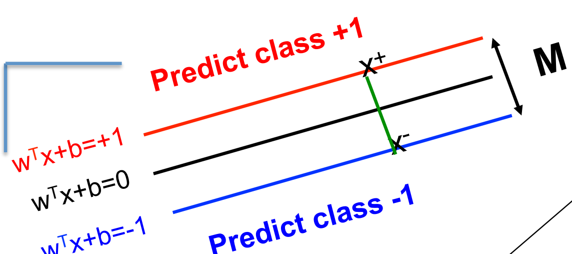
$-1 + \lambda w^T w = +1$

=>

$\lambda = 2/w^T w$

Dr. Yanjun Qi / UVA CS 6316 / f15

Putting it together



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

We can now define M in terms of w and b

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

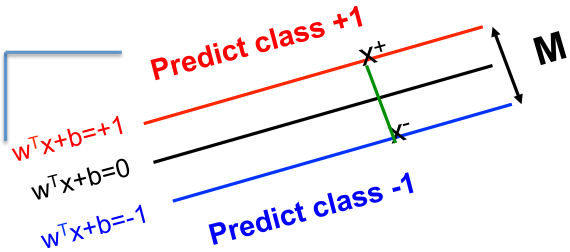
=>

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

=>

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

Finding the optimal parameters



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

We can now search for the optimal parameters by finding a solution that:

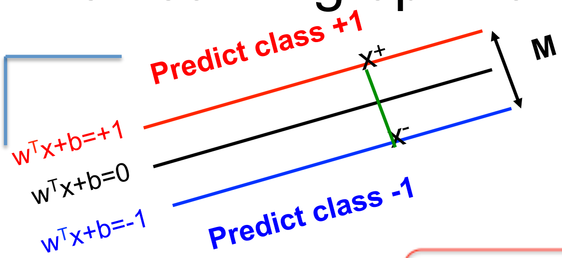
1. Correctly classifies all points
2. Maximizes the margin (or equivalently minimizes w^Tw)

Several optimization methods can be used: Gradient descent, simulated annealing, EM etc.

Today Recap

- ❑ Supervised Classification
- ❑ Support Vector Machine (SVM)
  - ✓ History of 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
  - ✓ Multiclass SVM

Optimization Step
i.e. learning optimal parameter for SVM



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

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

$w^Tx+b >= 1$

For all x in class - 1

$w^Tx+b <= -1$

A total of n constraints if we have n input samples

Support Vector Machine

Task	classification
Representation	Kernel Func K(xi, xj)
Score Function	Margin + Hinge Loss (optional)
Search/Optimization	QP with Dual form
Models, Parameters	Dual Weights

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

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

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

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

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

$$\sum_i \alpha_i y_i = 0, \qquad \alpha_i \geq 0 \qquad \forall i$$

References

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