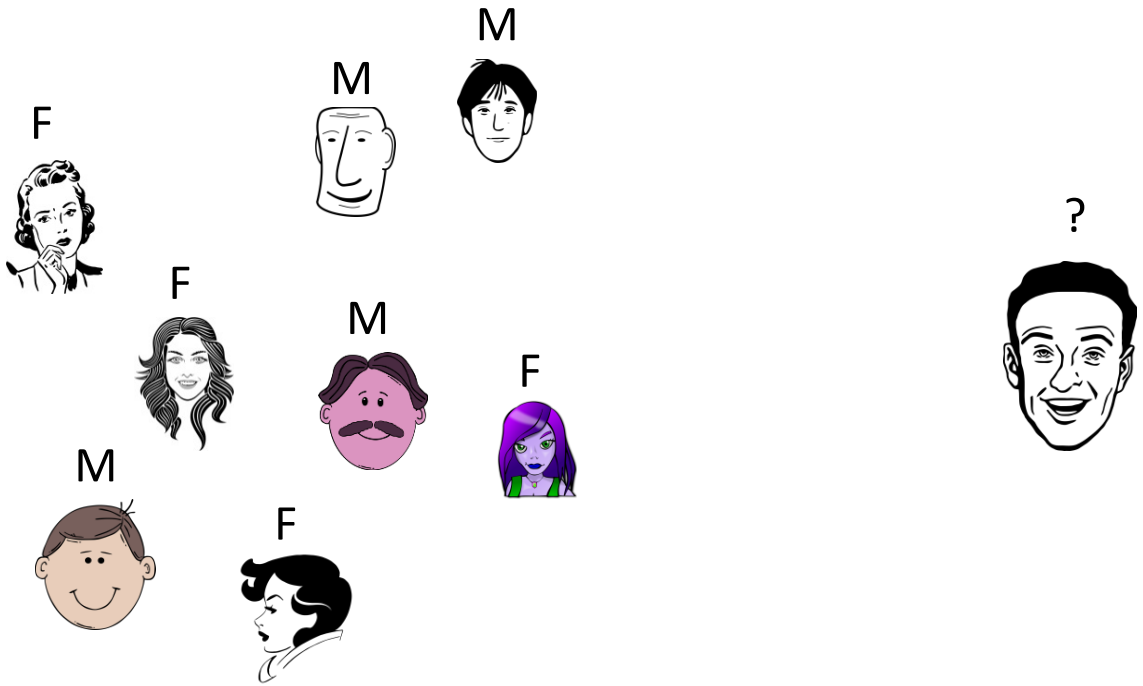
A Primer on Machine Learning, Classification, and Privacy

Anupam Datta

Fall 2016

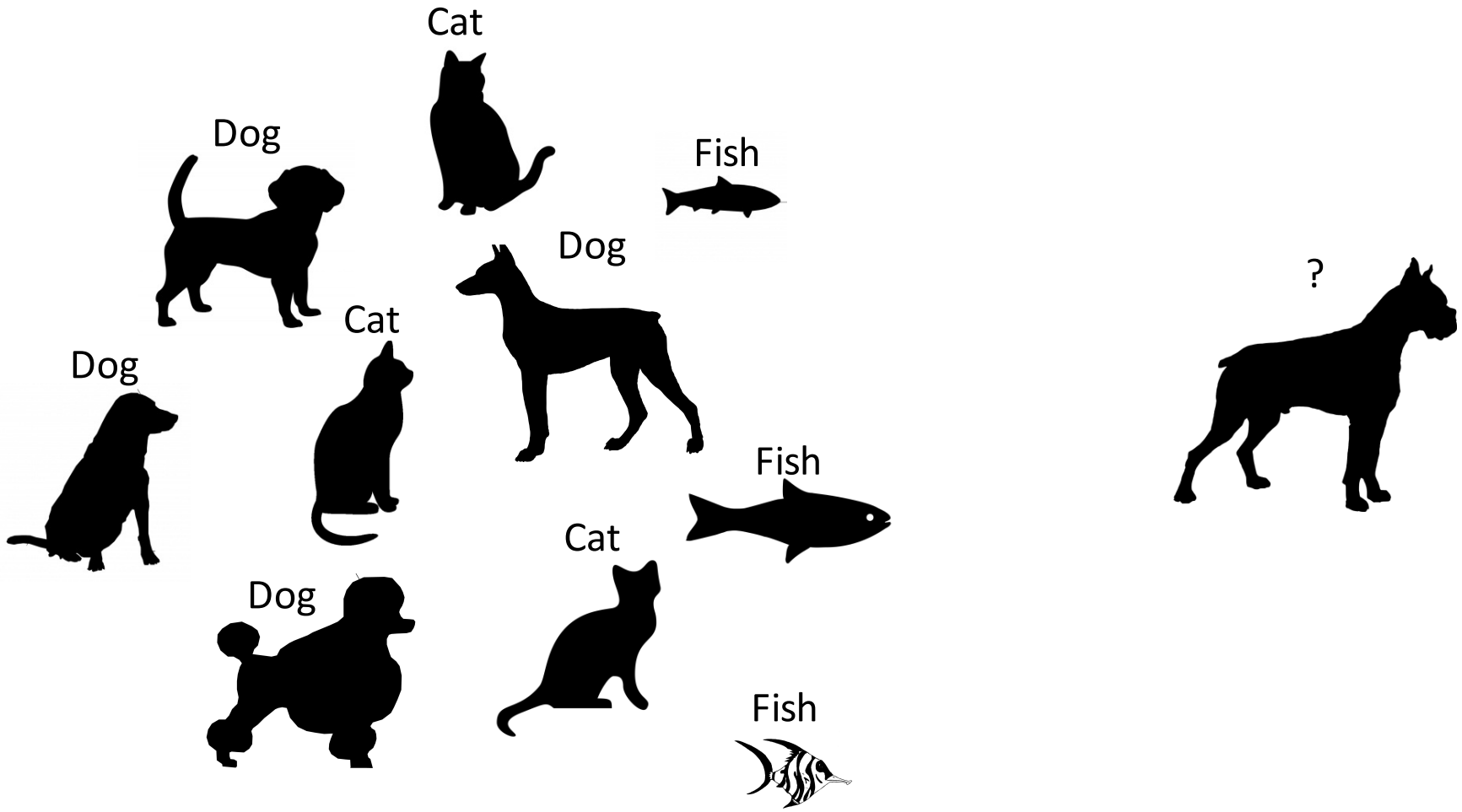
Machine Learning - Classification

- A *classification algorithm* takes as input some examples and their values (**labels**), and, using these examples, is able to predict the values of new data.



Machine Learning – Multiclass algorithms

- Multiclass algorithms: more than two labels



Machine Learning – Regression

- Labels are real values

Height (cm)	Weight (kg)	Age	Gender	Heart Rate (BPM)
183	78	35	M	72
155	60	48	F	66
160	88	67	F	90
…	…	…	…	…
149	55	25	M	?

ML is Everywhere!

- Leading paradigm in big data science (with application to MANY fields in CS/engineering).
- Infer user traits based on their online behavior (Google, Microsoft, Facebook, Amazon etc.)
- Image identification
- Face/Voice Recognition
- Natural Language Processing

Big Data for Advertising/Sales

amazon

Google

Expedia

bing

twitter

Anonymization and Consent

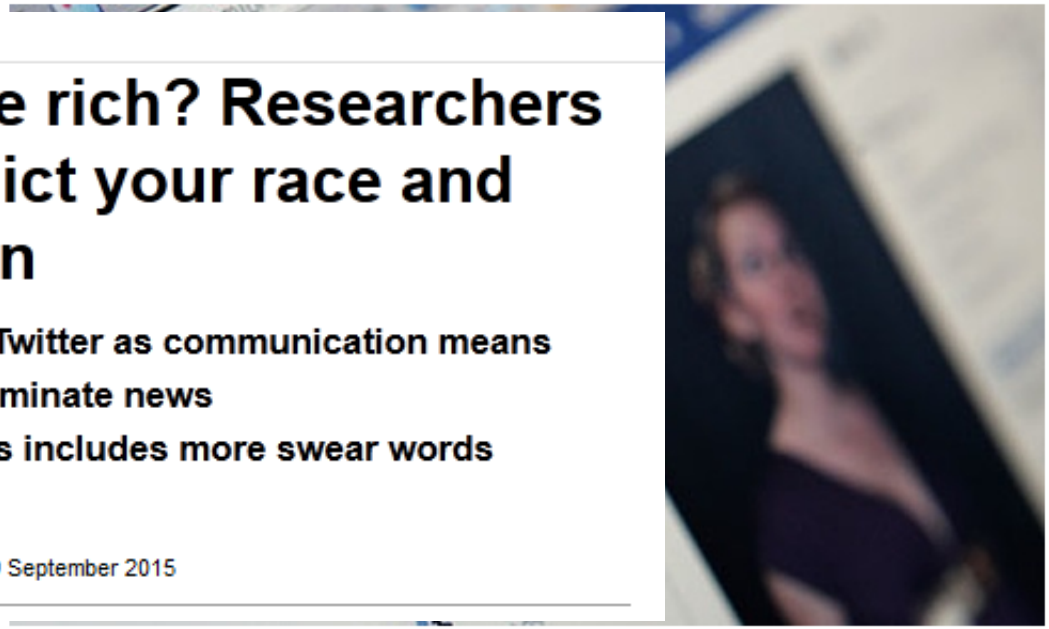
- Certain attributes are considered private/protected.
- Users must consent to their data being used.

…. But private/protected data can be inferred with frightening accuracy!

Raising Some Concerns…

Facebook users unwittingly revealing intimate secrets, study finds

Personal information including sexuality and drug use can be correctly inferred from public 'like' updates, according to study



Facebook: researchers were able to accurately infer a Facebook user's race, IQ, sexuality, substance use and political views using only their 'Likes' Photograph: Chris Jackson/Getty Images

Facebook users are unwittingly revealing intimate secrets - including their sexual orientation, drug use and political beliefs - using only public "like" updates, according to a study of online privacy.

Do YOU tweet like you're rich? Researchers find messages can predict your race and even how much you earn

- **Lower socioeconomic status users use Twitter as communication means**
- **High-income people use it more to disseminate news**
- **Text from those in lower income brackets includes more swear words**

By MARK PRIGG FOR DAILYMAIL.COM

PUBLISHED: 18:43 EST, 29 September 2015 | **UPDATED:** 20:42 EST, 29 September 2015

ML and Privacy

- Unfettered ML as a threat to privacy
  - Attacks on privacy using ML
  - Examining standard ML models to understand if they present a threat to privacy
  - Constraining standard ML to ensure that some privacy notions are respected

- ML as a tool for protecting privacy
  - Using ML to detect threats to privacy and related values

This Lecture

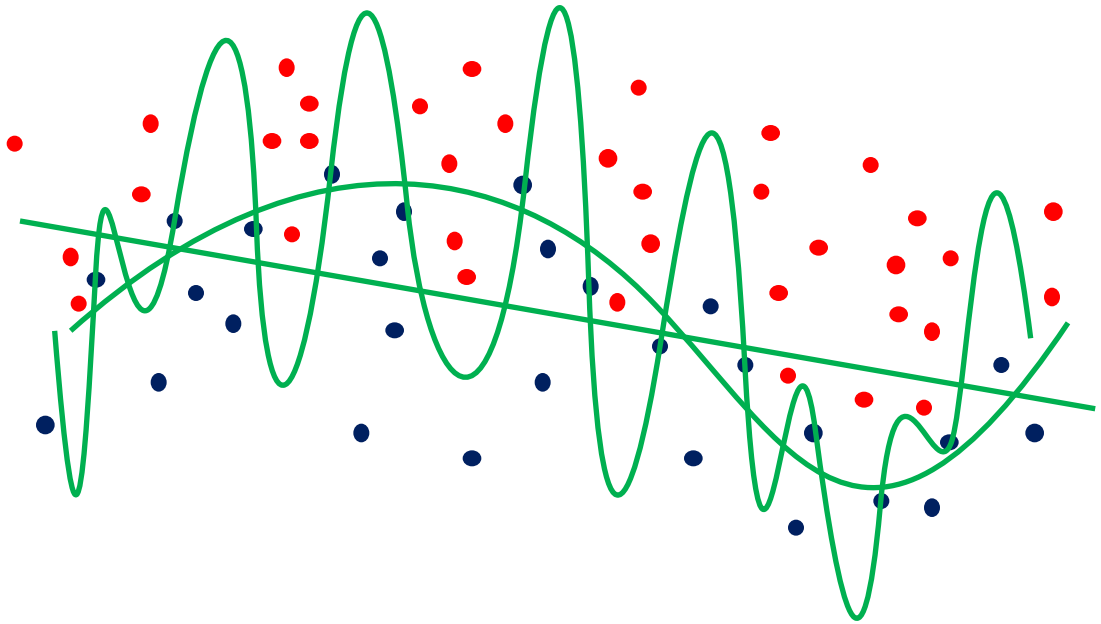
- Basic concepts in classification
- Some examples of classifiers
- Loss minimization
- Some case studies

Classification Algorithms

- We are given a dataset $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$
- assume that $\mathbf{x}_j \in \mathbb{R}^m$, m is the number of *features.*
- Each point $\mathbf{x}_i$ is assigned a value t_j in $\{-1,1\}$.

Learning a Classifier

- Our goal is to find a function $y: \mathbb{R}^m \rightarrow \{-1,1\}$ that fits the data: *what is the likeliest function to have generated our dataset?*

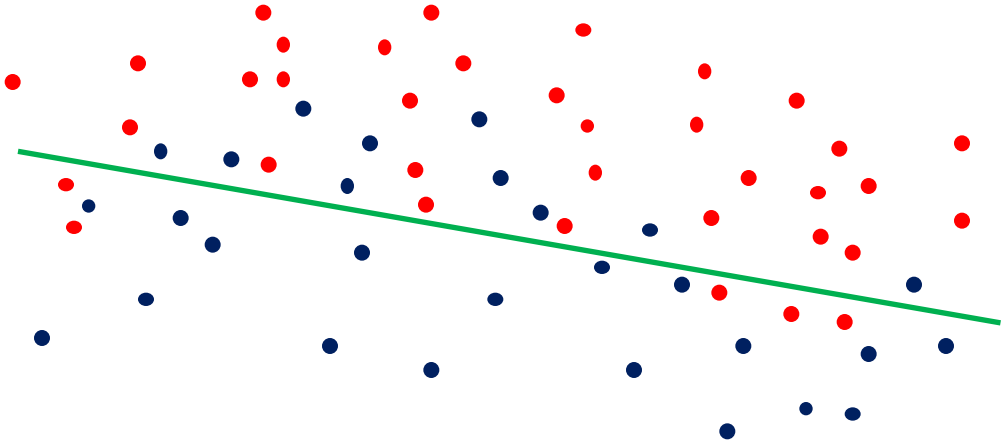


Testing a Classifier

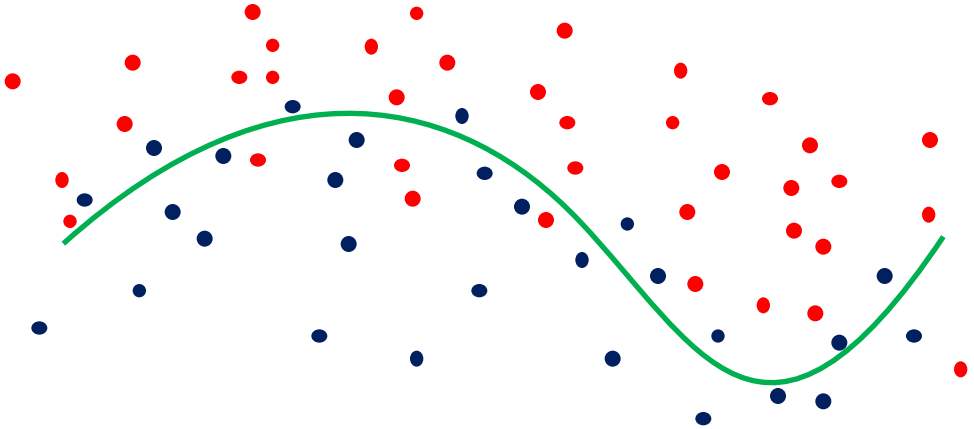
- A separate data set used to evaluate the prediction accuracy of the classifier.

- What percentage of the predictions are accurate on the test set?

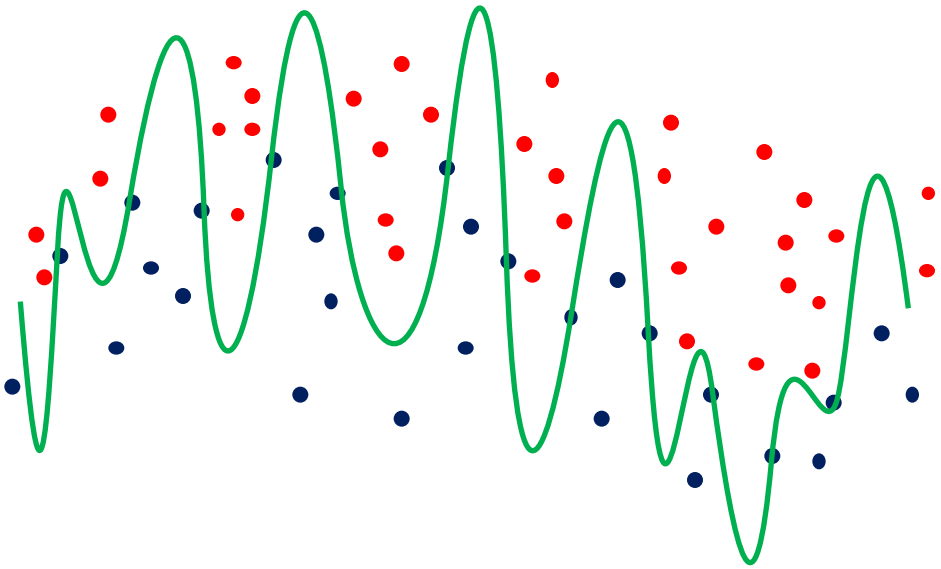
I.



II.



III.



Learning a Classifier

We are given a dataset as follows:

BMI	Overweight (1)	Not Overweight (-1)
32	2	0
30	5	0
28	4	3
26	2	3
24	2	1

Find a function $y: \mathbb{R}_+ \rightarrow \{-1,1\}$ that, given a BMI value x, predicts whether a person with this BMI is overweight.

Example adapted from A. Shashua, "Introduction to Machine Learning", ArXiv 2008

Outline

- Probability Basics
- Bayes Classifier
- Linear Classifier
- Support Vector Machines

Definition of Probability

- ***Experiment***: toss a coin twice
- ***Sample space***: possible outcomes of an experiment
  - S = {HH, HT, TH, TT}
- ***Event***: a subset of possible outcomes
  - A={HH}, B={HT, TH}, C={TT}
- ***Probability of an event*** : an number assigned to an event Pr(A)
  - Axiom 1: Pr(A) ≥ 0
  - Axiom 2: Pr(S) = 1
  - Axiom 3: For every sequence of disjoint events

$$\Pr(\bigcup_i A_i) = \sum_i \Pr(A_i)$$

Definition of Probability

- ***Experiment***: toss a coin twice
- ***Sample space***: possible outcomes of an experiment
  - S = {HH, HT, TH, TT}
- ***Event***: a subset of possible outcomes
  - A={HH}, B={HT, TH}, C={TT}

Assuming coin is fair,

P(A) = ¼ P(B) = ½ P(C) = ¼

What is the probability that we get at least one head?

P({HH, HT, TH}) = ¾

Joint Probability

- For events A and B, **joint probability** $Pr(A \cap B)$ stands for the probability that both events happen.
- Example: A={HH}, B={HT, TH}, what is the joint probability $Pr(A \cap B)$?
- $P(A \cap B) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Joint Probability

- ***Experiment***: toss a coin twice
- ***Sample space***: possible outcomes of an experiment
  - S = {HH, HT, TH, TT}
- ***Event***: a subset of possible outcomes
  - A={HH}, B={HT, TH}, C={TT}

- $P(A \cap B) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- $P(\{HH, HT, TH\})$

  $= P(\{HH\} \cup \{HT, TH\}) = P(A \cup B) = P(A)+P(B) - P(A \cap B)$

  $= \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$

Conditional Probability

- If A and B are events with $Pr(A) > 0$, the ***conditional probability of B given A*** is

$$Pr(B|A) = Pr(A \cap B)/Pr(A)$$

Example:
A={HH, TH}, B={HH}
Pr(B|A) =
Pr({HH})/Pr({HH, TH})
¼ / ½ = ½

Outline

- Probability Basics
- Bayes Classifier
- Linear Classifier
- Support Vector Machines

Bayesian Probability

Given two random variables X, Y, the conditional probability of $X = x$ given $Y = y$ is denoted $\Pr[\, X = x \mid Y = y \,]$:
"how likely is it that $X = x$, given that I know that $Y = y$?"

Let X be BMI, and Y be 1 if overweight, -1 otherwise.

More generally, X is the distribution over datapoints, and Y is the distribution of values.

$\Pr[Y \mid X]$ is called the *posterior distribution*

$\Pr[X \mid Y]$ is called the *class conditional likelihood*

Bayesian Probability

BMI	Overweight (1)	Not Overweight (-1)
32	2	0
30	5	0
28	4	3
26	2	3
24	2	1

$$\Pr[X = x \mid Y = y] = \frac{\Pr[X = x, Y = y]}{\Pr[Y = y]}$$

$$\Pr[X = 30 \mid Y = 1] = \frac{5/22}{15/22} = \frac{5}{15} = \frac{1}{3}$$

Example adapted from A. Shashua, "Introduction to Machine Learning", ArXiv 2008

Bayesian Probability

Given a classifier $y: \mathbb{R}^n \to \{-1,1\}$, how likely is y to misclassify a data point?

$$\begin{aligned}\Pr[\text{error}] &= \Pr_{\mathbf{x}\sim X}[y(\mathbf{x}) \neq Y] \\ &= \Pr_{\mathbf{x}\sim X}[y(\mathbf{x}) = 1, Y = -1] \\ &+ \Pr_{\mathbf{x}\sim X}[y(\mathbf{x}) = -1, Y = 1]\end{aligned}$$

Bayesian Probability

Maximizing the posterior probability:

BMI	Overweight (1)	Not Overweight (-1)
32	2	0
30	5	0
28	4	3
26	2	3
24	2	1

BMI	$p(Y = 1 \mid X = \mathbf{x})$	$p(Y = -1 \mid X = \mathbf{x})$
32	1	0
30	1	0
28	$4/7$	$3/7$
26	$2/5$	$3/5$
24	$2/3$	$1/3$

Bayesian Probability

So, a classifier that minimizes error probability could be of the form

$$y(x) = \begin{cases} 1 \text{ if } x \geq y_0 \\ -1 \text{ otherwise} \end{cases}$$

Where $26 < y_0 \leq 28$

In general:

$y(\mathbf{x})$ should be 1 iff

$$p(Y = 1 \mid \mathbf{x}) \geq p(Y = -1 \mid \mathbf{x})$$

Outline

- Probability Basics
- Bayes Classifier
- Linear Classifier
- Support Vector Machines

Learning a discriminant function

Assume that the data was generated using an (unknown) function $f: \mathbb{R}^n \to \{-1,1\}$, perhaps with some error.

Objective

Assuming that f belongs to some class of functions $\mathcal{C}$, find a function $y \in \mathcal{C}$ that estimates f well.

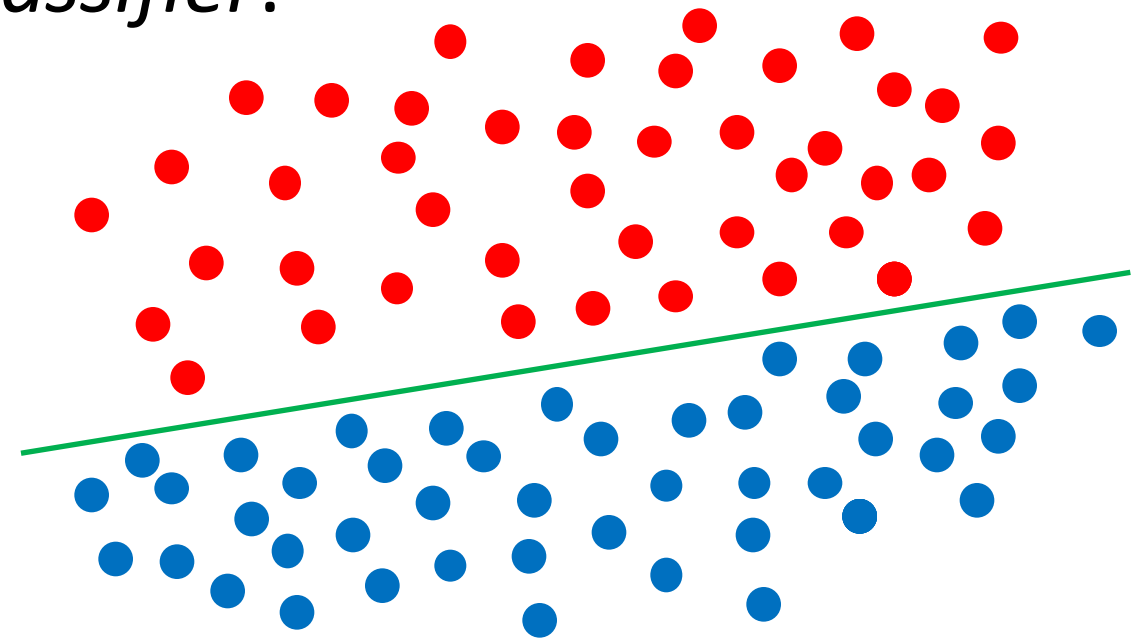
This function is called a *discriminant*.

Linear Classifiers

Assumption: data was generated by some function $f(\mathbf{x}) = \begin{cases} 1 \text{ if } \mathbf{w}^T\mathbf{x} + w_0 \geq 0 \\ -1 \text{ otherwise.} \end{cases}$

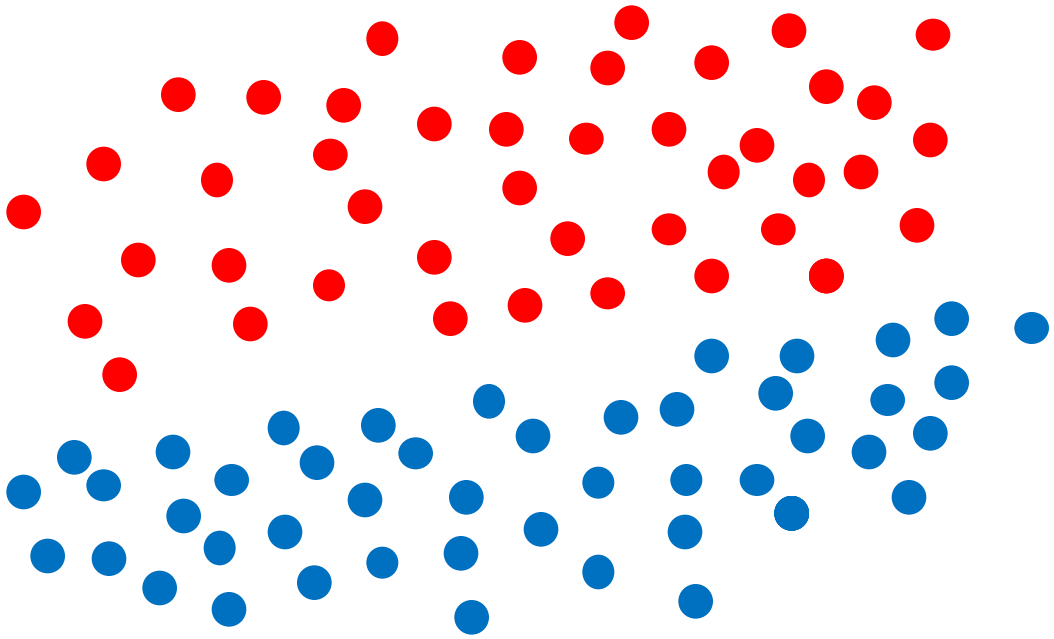
f is called a *linear classifier*.

$\boldsymbol{f(x, y) = \mathbb{I}(w_1 x + w_2 y + w_0 \geq 0)}$



Linear Classifiers

Question: given a dataset, how would we determine which linear classifier is “good”?



Least Squared Error

Let $y(\mathbf{w}, \mathbf{x}) = \mathbb{I}(\mathbf{w}^T \mathbf{x} \geq 0)$

Objective:

$$\min_{\mathbf{w} \in \mathbb{R}^m, w_0} \sum_j \left(y(\mathbf{w}, \mathbf{x}_j) - t_j \right)^2$$

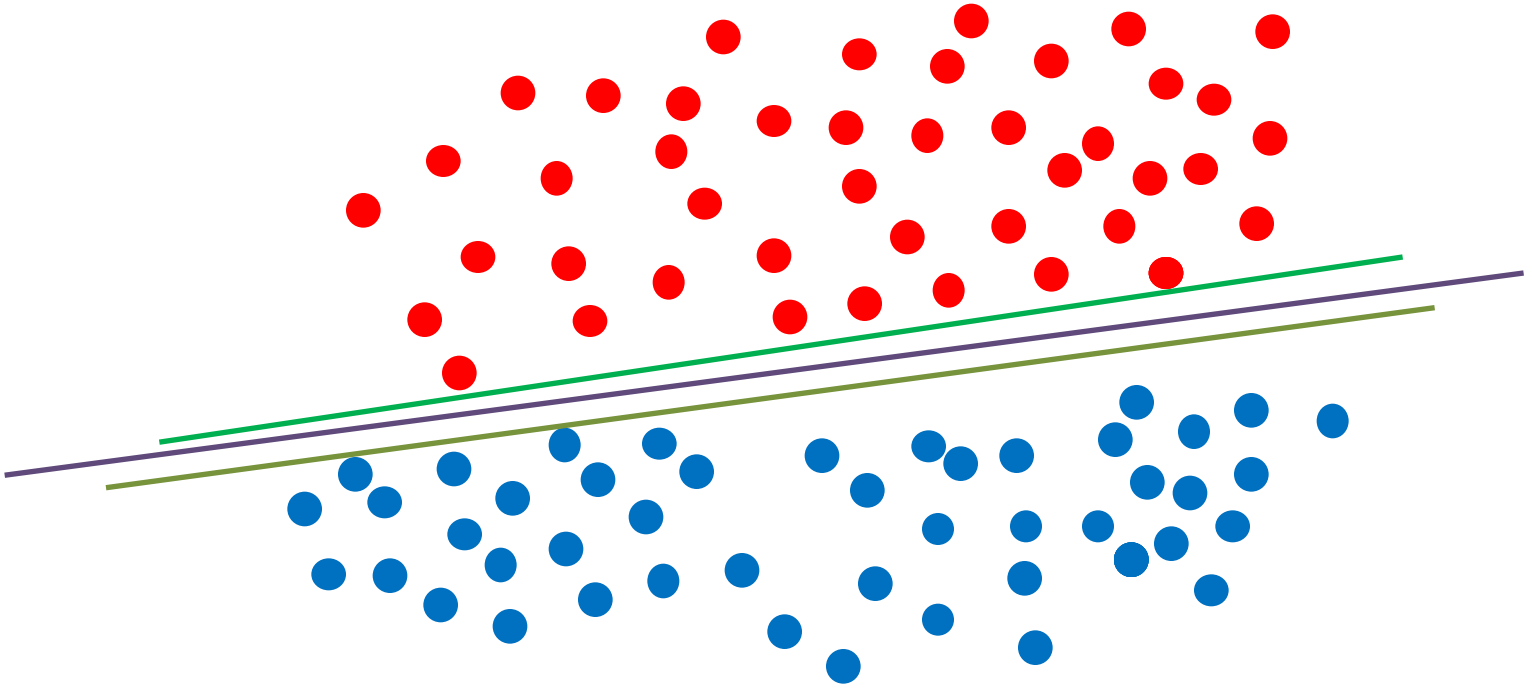
"minimize the distance between the outputted labels and the actual labels"

Other loss functions are possible!

Outline

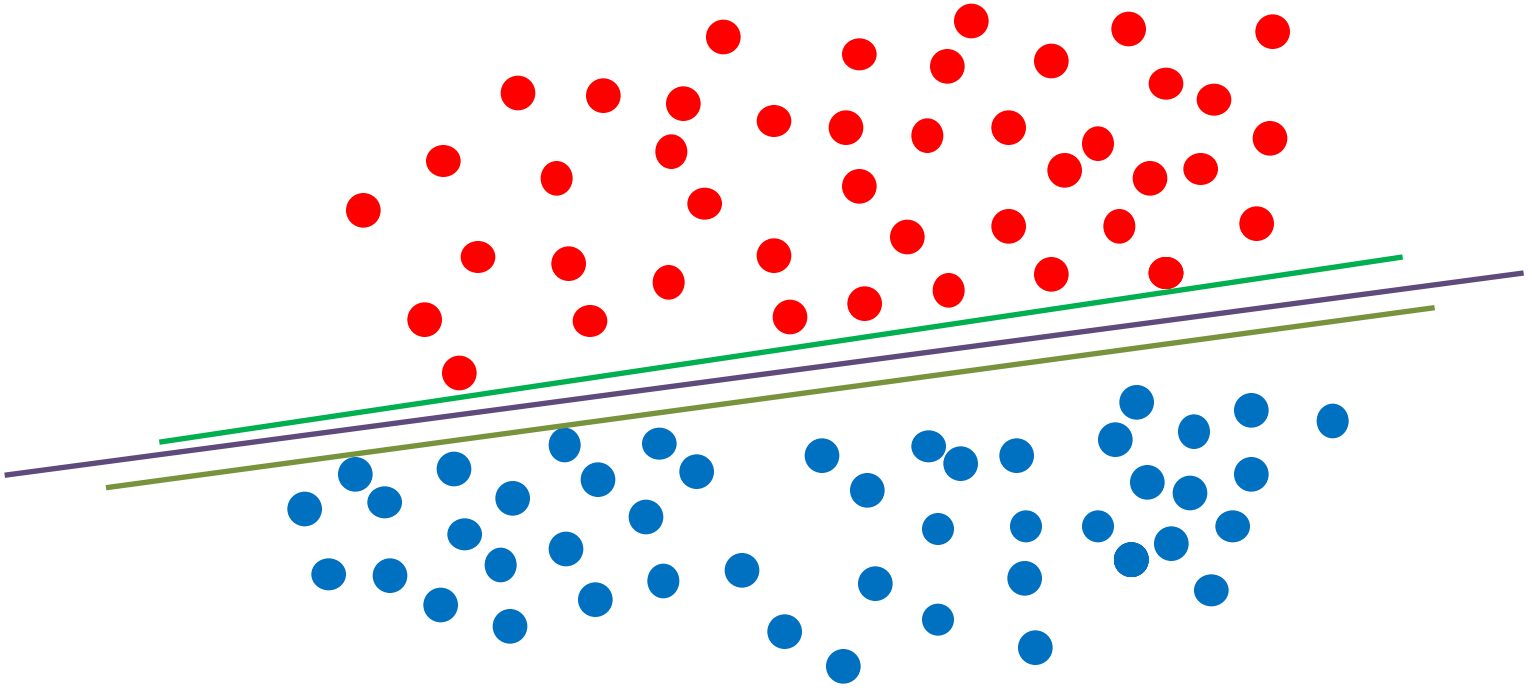
- Probability Basics
- Bayes Classifier
- Linear Classifier
- Support Vector Machines

Support Vector Machines



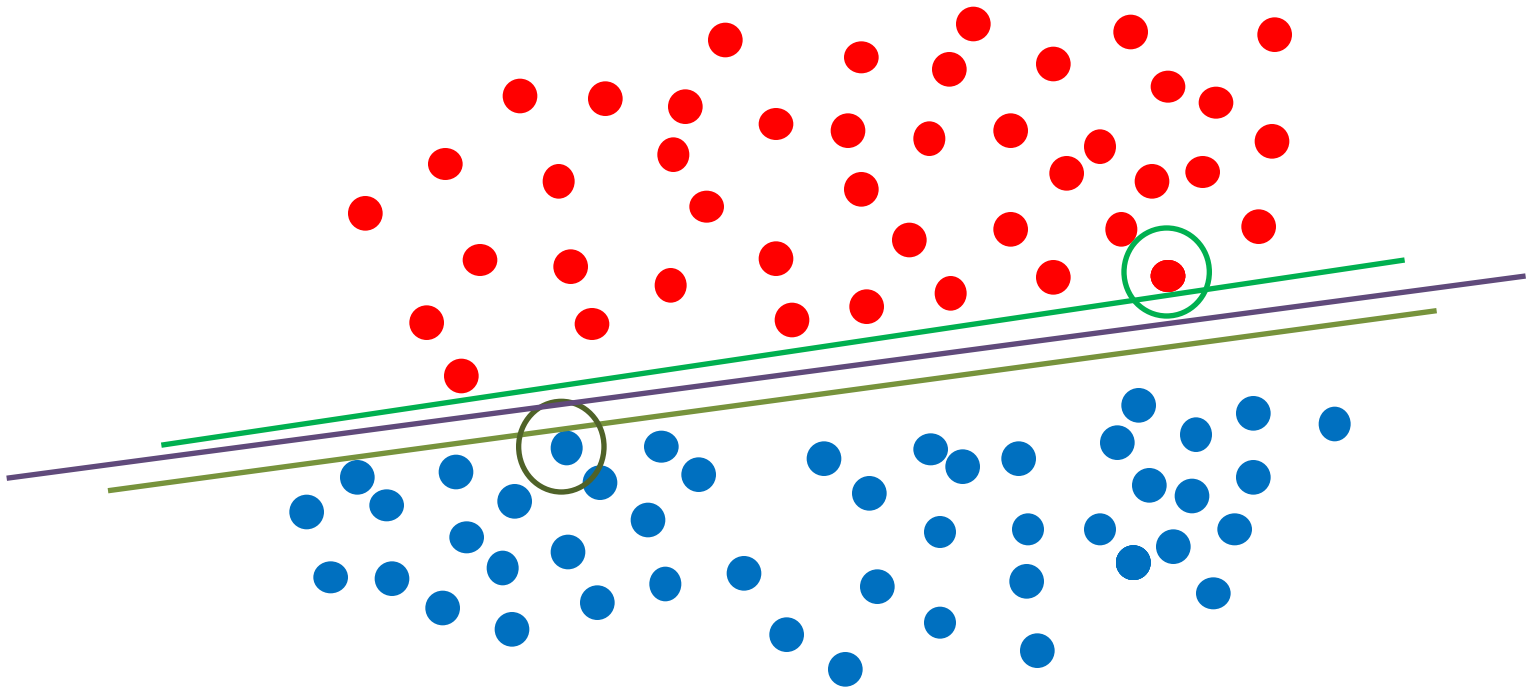
Many candidates for a linear function minimizing LSE; which one should we pick?

Support Vector Machines



Question: why is the purple line “good”?

Support Vector Machines: Key Insight



One possible approach:
find a hyperplane that is “perturbation resistant”

Support Vector Machines: Key Insight

- How to mathematical represent this idea?
- Margin: Minimal distance from points to classifier hyperplane
- Idea: Pick hyperplane that maximizes margin
- Can we formulated as optimization problem

Support Vector Machines

- The distance between the hyperplane $\mathbf{w}^T\mathbf{x} + w_0 = 0$ and some $\mathbf{x}_0 \in \mathbb{R}^m$ is $\frac{|\mathbf{w}^T\mathbf{x}_0 + w_0|}{\|\mathbf{w}\|}$
- Given a dataset $\mathbf{x}_1, \ldots, \mathbf{x}_n$ labeled $t_1, \ldots, t_n$, the *margin* of $(\mathbf{w}, w_0)$ with respect to the dataset is the minimal distance between the hyperplane defined by $(\mathbf{w}, w_0)$ and the datapoints.

Support Vector Machines

To find the best hyperplane, solve:

$$\max_{\mathbf{w},w_0} \left\{ \frac{1}{\|\mathbf{w}\|} \min_j t_j(\mathbf{w}^T \mathbf{x}_j + w_0) \right\}$$

Subject to

$$t_j\left(\mathbf{w}^T \mathbf{x}_j + w_0\right) \geq 0, \forall j$$

It can be shown that this is equivalent to minimizing $\|\mathbf{w}\|$ subject to the constraints.

Acknowledgments

- Augmented slides from Yair Zick for Fall 2015 version of 18734
- Material Adapted from:
  - C.M. Bishop, “Pattern Recognition & Machine Learning”, Springer, 2006
  - A. Shashua, “Introduction to Machine Learning – 67577”, Fall 2008 Lecture Notes, Arxiv.

Bayesian Probability

The expression

$$\int_{\mathbf{x}:y(\mathbf{x})=1} p(\mathbf{x},-1)\partial\mathbf{x} + \int_{\mathbf{x}:y(\mathbf{x})=-1} p(\mathbf{x},1)\partial\mathbf{x}$$

is minimized if we assign $\mathbf{x}$ a label $y \in \{-1,1\}$ for which the expression $p(\mathbf{x}, y)$ is maximized;

i.e. $y(\mathbf{x})$ should be 1 iff $p(\mathbf{x},1) \geq p(\mathbf{x},-1)$

$p(\mathbf{x},1) \geq p(\mathbf{x},-1) \leftrightarrow$

$p(Y = 1 \mid \mathbf{x})p(\mathbf{x}) \geq p(Y = -1 \mid \mathbf{x})p(\mathbf{x}) \leftrightarrow$

$p(Y = 1 \mid \mathbf{x}) \geq p(Y = -1 \mid \mathbf{x})$

Bayesian Probability

Given a classifier $y: \mathbb{R}^n \to \{-1,1\}$, how likely is y to misclassify a data point?

Shorthand:

$$\Pr[X = \mathbf{x}, Y = y] = p(\mathbf{x}, y)$$
$$\Pr[X = \mathbf{x}] = p(\mathbf{x})$$
$$\Pr[Y = y] = p(y)$$

$$\Pr[\text{error}] = \Pr_{\mathbf{x}\sim X}[y(\mathbf{x}) \neq Y]$$
$$= \Pr_{\mathbf{x}\sim X}[y(\mathbf{x}) = 1, Y = -1] + \Pr_{\mathbf{x}\sim X}[y(\mathbf{x}) = -1, Y = 1]$$
$$= \int_{\mathbf{x}:y(\mathbf{x})=1} p(\mathbf{x}, -1)\partial\mathbf{x} + \int_{\mathbf{x}:y(\mathbf{x})=-1} p(\mathbf{x}, 1)\partial\mathbf{x}$$

Bayesian Inference

Three approaches:

a) Estimate $p(\mathbf{x}, y)$, and use it to get a posterior distribution, from which we assign values minimizing error.

b) Infer $p(y \mid \mathbf{x})$, and use it to obtain an estimate.

c) Infer a function $y: \mathbb{R}^m \to \{-1,1\}$ directly

Least Squared Error

If we assume that $t = y(\mathbf{w}, \mathbf{x}) + \varepsilon$, where ε is noise generated by a Gaussian distribution, then minimizing LSE is equivalent to *maximum likelihood estimation* (MLE)

$$\max_{\mathbf{w} \in \mathbb{R}^m, w_0} \Pr[\, t_1, \ldots, t_n \mid \mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{w}, w_0 \,]$$

"How likely are the observed labels, given the dataset and the chosen classifier?"