Machine Learning

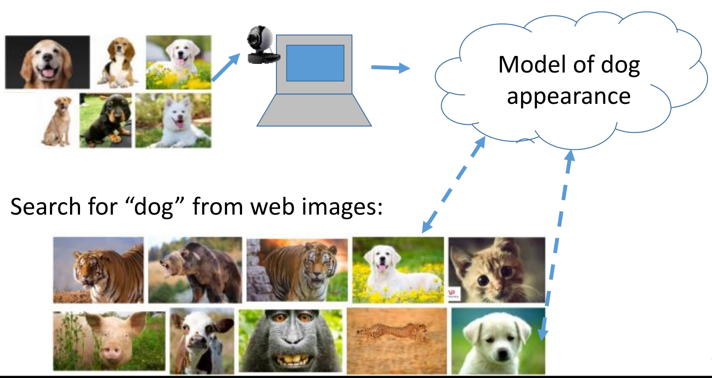
CISC 5800

Dr Daniel Leeds

What is machine learning

- Finding patterns in data
- Adapting program behavior

Dog photos and the internet



Change radio channel when user says "change channel"

- Distinguish user's voice from music
- Understand what user has said

What's covered in this class

- Theory: describing patterns in data
  - Probability
  - Linear algebra
  - Calculus/optimization
- Implementation: programming to find and react to patterns in data
  - Popular and successful algorithms
  - Python
  - Data sets of text, speech, pictures, user actions, neural data…

6

Outline of topics

- Groundwork: probability and slopes
- Classification overview: Training, testing, and overfitting
- Basic classifiers: Naïve Bayes and Logistic Regression
- Advanced classifiers: Neural networks and support vector machines

  Deep learning **Kernel methods**
- Dimensionality reduction: Feature selection, information criteria
- Graphical models: Hidden Markov Model
- Expectation-Maximization
- Learning theory

7

What you need to do in this class

- Class attendance
- Assignments: homeworks (4-5) and final project
- Exams: midterm and final
- Don't cheat
  - You may discuss course topics with other students, but your submitted work must be your own. Copying is not allowed.

8

Resources

- Office hours: Wednesday 4-5pm and by appointment LL 610H

  Teaching Assistant: TBA LL 6th floor
- Course web site: http://storm.cis.fordham.edu/leeds/cisc5800
- Fellow students
- Textbooks/online notes
- Python

python

Andrew Ng's Stanford course notes

CS229 Machine Learning Autumn 2016

9

Probability and basic calculus

Probability and basic calculus

Probability

What is the probability that a child likes chocolate?

- Ask 100 children
- Count who likes chocolate
- Divide by number of children asked

Name	Chocolate?
Sarah	Yes
Melissa	Yes
Darren	No
Stacy	Yes
Brian	No

P("child likes chocolate") = $\frac{85}{100} = 0.85$

In short: P(C=true)=0.85 C="child likes chocolate"

General probability properties

P(A) means "Probability that statement A is true"

- $0 \leq \text{Prob}(A) \leq 1$
- Prob(True)=1
- Prob(False)=0

Random variables

A variable can take on a value from a given set of values:

- {True, False}
- {Cat, Dog, Horse, Cow}
- {0,1,2,3,4,5,6,7}

A random variable holds each value with a given probability

Example: **binary variable** LikesChocolate

- P(LikesChocolate) = P(LikesChocolate=True) = 0.85

14

Complements

C="child likes chocolate"

$P(C{=}true) = \frac{85}{100} = 0.85$

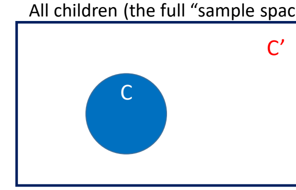
What is the probability that a child DOES NOT like chocolate?

Complement: "C=False" <-> "child doesn't like chocolate"

P(C=false) = **.15**

In general: P(A=false) = **1-P(A)**

All children (the full "sample space")



15

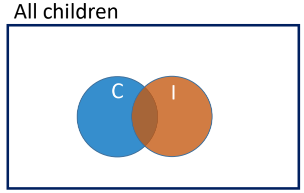
Addition rule

Prob(A=tr or B=tr) = ???

Name	Chocolate?	Ice cream?
Sarah	Yes	No
Melissa	Yes	Yes
Darren	No	No
Stacy	Yes	Yes
Brian	No	Yes

C="child likes chocolate"
I="child likes ice cream"

All children



17

Joint probabilities

C="child likes chocolate"
I="child likes ice cream"

Across 100 children:

- 55 like chocolate AND ice cream — P(I=True, C=True)**=.55**
- 30 like chocolate but not ice cream
- 5 like ice cream but not chocolate
- 10 don't like chocolate nor ice cream

P(I=False, C=True) **=.3**
P(I=True,C=False) **=.05**

P(I=True) =.6
P(C=True) =.85

19

Conditional probability

Also, **Multiplication Rule:**

P(A,B) = P(A|B) P(B)

P(A,B):Probability A and B are both true

Across 100 children:
- 55 like chocolate AND ice cream P(C=t,I=t)**=0.55**
- 30 like chocolate but not ice cream P(C=t,I=f) **=0.3**
- 5 like ice cream but not chocolate P(C=f,I=t) **=0.05**
- 10 don't like chocolate nor ice cream P(C=f,I=f) **=0.1**

- Prob(C|I) : Probability child likes chocolate given s/he likes ice cream

$$P(C|I) = \frac{P(C,I)}{P(I)} = \frac{P(C,I)}{P(C=true,I)+P(C=false,I)}$$

21

Marginal and conditional probabilities

For two **binary** random variables A and B
- P(A) = P(A, B=True) + P(A, B=False)
- P(B) = P(A=True,B)+P(A=false,B)

For **marginal probability** P(X), "marginalize" over all possible values of the other random variables

- Prob(C|I) : Probability child likes chocolate given s/he likes ice cream

$$P(C|I) = \frac{P(C,I)}{P(I)} = \frac{P(C,I)}{P(C=true,I)+P(C=false,I)}$$

22

Independence

If the truth value of B does not affect the truth value of A, we say A and B are **independent**.

- P(A|B) = P(A)

- P(A,B) = P(A) P(B)

23

Multi-valued random variables

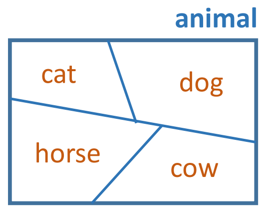
A random variable can hold more than two values, each with a given probability
- P(Animal=Cat)=0.5
- P(Animal=Dog)=0.3
- P(Animal=Horse)=0.1
- P(Animal=Cow)=0.1

24

Probability rules: multi-valued variables

For given random variable A:

- P($A = a_i$ and $A = a_j$) = 0 if $i \neq j$
- $\sum_i P(A = a_i) = 1$
- $P(A = a_i) = \sum_j P(A = a_i, B = b_j)$



a is a value assignment for variable A

25

Probability table

- P(G=C,H=True)**=0.15**
- P(H=True) **=0.75**
- P(G=C|H=True) $=\frac{.15}{.75}$**=0.2**
- P(H=True|G=C) $=\frac{.15}{.2}$**=0.75**

Grade	Honor-Student	P(G,H)
A	False	0.05
B	False	0.05
C	False	0.05
D	False	0.1
A	True	0.3
B	True	0.2
C	True	0.15
D	True	0.1

26

Bayes rule

$$\mathrm{P(A|B)} = \frac{P(B|A)P(A)}{P(B)}$$

Terminology:

- P(A|B) is the "posterior probability"
- P(B|A) is the "likelihood"
- P(A) is the "prior probability"

We will spend (much) more time with Bayes rule in following lectures

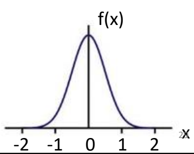
28

Continuous random variables

A random variable can take on a continuous range of values

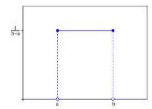
- From 0 to 1
- From 0 to ∞
- From $-\infty$ to ∞

Probability expressed through a
"probability density function" **f(x)**

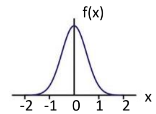


Common probability distributions

- Uniform: $f_{uniform}(x) = \begin{cases} \frac{1}{b-a} & if\ a \leq x \leq b \\ 0 & otherwise \end{cases}$



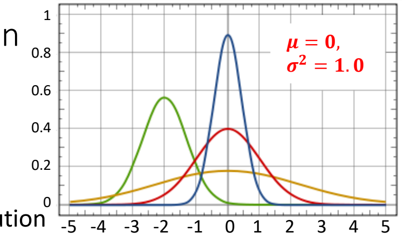
- Gaussian: $f_{gauss}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$



30

The Gaussian function

$f_{gauss}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$



- Mean μ – center of distribution
- Standard deviation σ – width of distribution

- Which color is μ=-2, σ^2=0.5? Which color is μ=0, σ^2=0.2?

- $N(\mu_1, \sigma_1^2) + N(\mu_2, \sigma_2^2) = N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

32

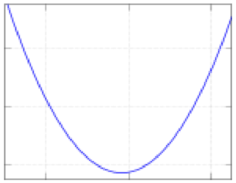
Probability and basic calculus

33

Calculus: finding the slope of a function

What is the minimum value of: f(x)=x^2-5x+6

Find value of x where slope is 0



General rules: slope of f(x): $\frac{d}{dx}f(x) = f'(x)$

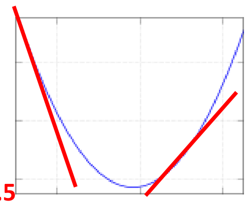
- $\frac{d}{dx}x^a = ax^{a-1}$
- $\frac{d}{dx}kf(x) = kf'(x)$
- $\frac{d}{dx}[f(x) + g(x)] = f'(x) + g'(x)$

34

Calculus: finding the slope of a function

What is the minimum value of: $f(x)=x^2-5x+6$

- f'(x)=**2x-5**
- What is the slope at x=5? **f'(5)=5**
- What is the slope at x=-3? **f'(-3)=-11**



- What value of x gives slope of 0? **x=2.5**

36

More on derivatives: $\frac{d}{dx}f(x) = f'(x)$

- $\frac{d}{dx}f(w) = 0$ -- w is not related to x, so derivative is 0
- $\frac{d}{dx}\left(f(g(x))\right)=g'(x) \cdot f'(g(x))$
- $\frac{d}{dx}\log x = \frac{1}{x}$
- $\frac{d}{dx}e^x = e^x$

37

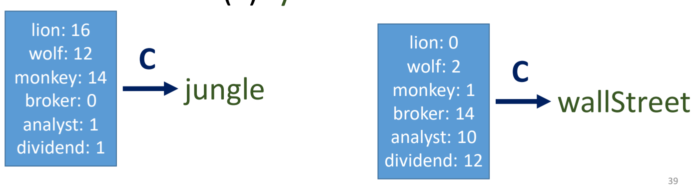
Review of classifiers

38

The goal of a classifier

- Learn function C to maximize
  correct labels (Y) based on features (X)

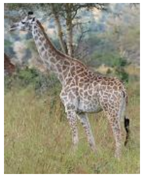
C(x)=y



39

Giraffe detector

- Label x : height
- Class y : True or False ("is giraffe" or "is not giraffe")



Learn optimal classification parameter(s)

- Parameter: x^{thresh}

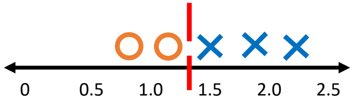
Example function:

$$C(x) = \begin{cases} True & \text{if } x > x^{thresh} \\ False & \text{otherwise} \end{cases}$$

40

Learning our classifier parameter(s)

- Adjust parameter(s) based on observed data
- Training set: contains features and corresponding labels



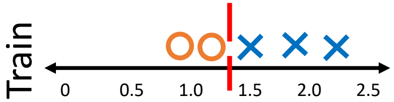
X	Y
1.5	True
2.2	True
1.8	True
1.2	False
0.9	False

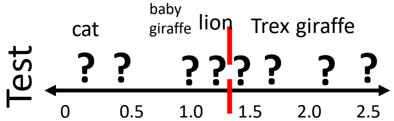
41

The testing set

Testing set must be distinct from training set!

- Does classifier correctly label new data?





Example "good" performance:
90% correct labels

42

Be careful with your training set

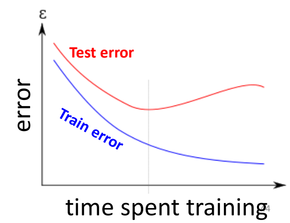
- What if we train with only baby giraffes and ants?
- What if we train with only T rexes and adult giraffes?

43

Training vs. testing

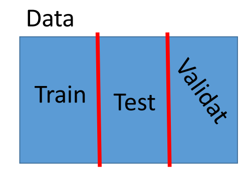
- **Training**: learn parameters from set of data in each class
- **Testing**: measure how often classifier correctly identifies new data

- More training reduces classifier error ε
- Too much training ~~data~~ iterations causes worse testing error – overfitting



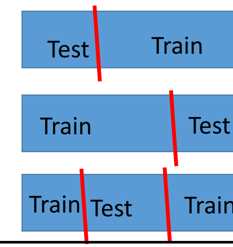
Dividing data sets

Three way divide: Train / test / validate



Cross-validation

- k-fold
- Leave-one-out



45

What is "good" classifier performance?

How well can you do if:

- You guess randomly?
  - If there are 2 equally likely classes: 0.5 accuracy
  - If there are k classes: 1/k accuracy
- You guess the most-common class?
  - If most common class has P(y) probability: P(y) accuracy
  - Wise to check accuracy for each class separately

47