ECE-175A

Elements of Machine Intelligence - I

Nuno Vasconcelos
(Ken Kreutz-Delgado)

ECE Department, UCSD

The course

- The course will cover basic, but important, aspects of machine learning and pattern recognition

- We will cover a lot of ground, at the end of the quarter you'll know how to implement a lot of things that may seem very complicated today

Logistics

- Exams: 1 mid-term - 35%
  1 final – 45% (covers everything)

- Quizzes (20%):
  - one problem every week.
  - a small computational problem. By small, I mean in terms of concepts, thinking, etc.
  - some computational problems will require a fair amount of computer power, e.g. a few hours on a laptop.
  - be sure to start early
  - will count 20%, but almost impossible without it.
  - will give you the hands-on experience needed to be able to claim that you really know learning!

Quiz policy

- Homework sets
  - include pen+paper and computer problems
  - pen+paper are not due, for practice only
- Quizzes: computer problems are due
  - issue and due dates specified on course web site
  - it is your responsibility to keep track of the dates
  - Quizzes are individual
  - due on the specified due-date. No exceptions, unless there is an extension to the entire class.
- The final course grade will be curve-based.
  - do not get discouraged if a homework problem is hard, just put some extra effort into solving it. Try your best.

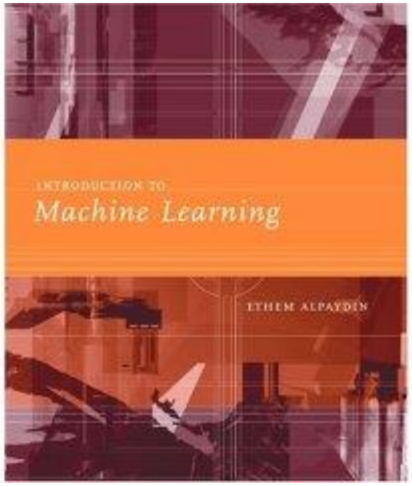
Academic integrity

- not allowed to
  - talk to friends or classmates about graded problems
  - use **any** ECE175A material not explicitly handed out by us
  - this includes consulting **any** websites other than the class web site, piazza, or canvas
- graded problems
  - think of these as part of an exam
  - don't ask questions you would not ask on exam room
  - don't ask "does my result look OK?," or "did anyone get something like this?"
- all work done for grade must be **individual**
  - we refer violations to UCSD Academic Integrity Office
  - despite these warnings there are always 3-4 cases a year

Resources

- Course web page:
  http://www.svcl.ucsd.edu/~courses/ece175/
  - all materials, except homework and exam solutions will be available there.
- Discussion forum:
  - We will be using piazza. You will get email.
- Course Instructor:
  - Nuno Vasconcelos, nuno@ece.ucsd.edu, EBU 1- 5602
  - office hours: TBA

Texts



- Required:
  - "Introduction to Machine Learning"
  - Ethem Alpaydin, MIT Press, 2004

- Various other good, but optional, texts:
  - "Pattern Classification", Duda, Hart, Stork, Wiley, 2001
  - "Elements of Statistical Learning", Hastie, Tibshirani, Fredman, 2001
  - "Pattern Recognition and Machine Learning", C.M. Bishop, Springer, 2007.

- Prerequisites you must know well:
  - "Linear Algebra", Gilbert Strang, 1988
  - "Fundamentals of Applied Probability", Drake, McGraw-Hill, 1967

Why Machine Learning?

- Good question! After all, many systems & processes in the world that are well-modeled by deterministic equations
  - E.g. f = m a; V = I R, Maxwell's equations, and other physical laws.
  - There are acceptable levels of "noise", "error", and other "variability".
  - In such domains, we don't need statistical learning.
- However, learning is necessary when there is a need for predictions about, or classification of, poorly known and/or random vector data Y, that
  - represents important events, situations, or objects in the world;
  - which may (or may not) depend on other factors (variables) X;
  - is impossible or too difficult to derive an exact, deterministic model for;

Examples and Perspectives

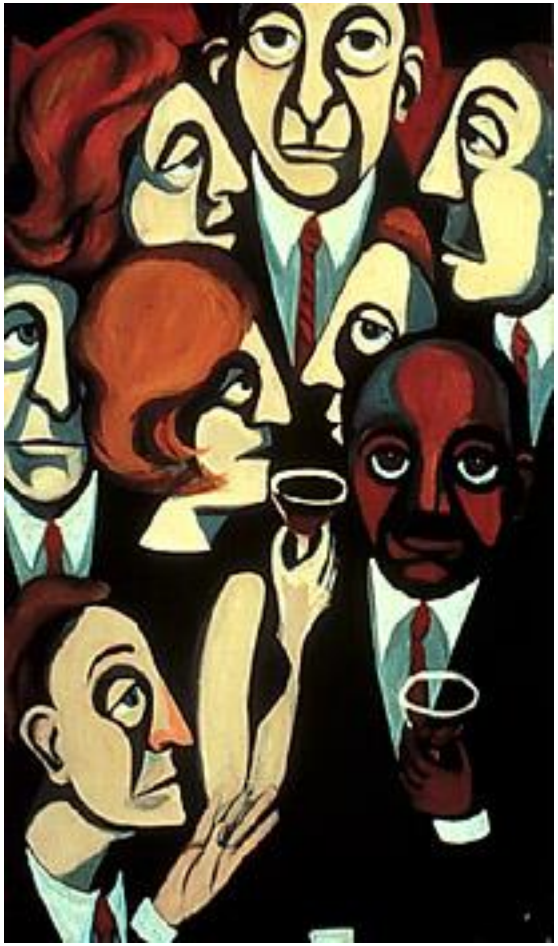
- The "Data-Mining" viewpoint:
  - huge amounts of data that **does not follow deterministic rules**
  - E.g. given an history of thousands of customer records and some questions that I can ask you, how do I predict that you will pay on time?
  - Impossible to derive a theory for this, must be learned

While many associate learning with data-mining, it is by no means the only important application or viewpoint.

- The Signal Processing viewpoint:
  - Signals combine in ways that depend on **"hidden structure"** (e.g. speech waveforms depend on language, grammar, etc.)
  - Signals are usually subject to significant amounts of **"noise"** (which sometimes means "things we do not know how to model")

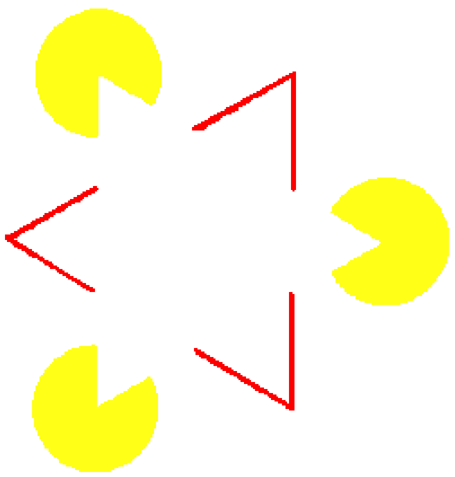
Examples - Continued

- Signal Processing viewpoint (Cont'd)
  - E.g. the Cocktail Party Problem:
    - Although there are all these people talking loudly at once, you can still understand what your friend is saying.
    - How could you build a chip to separate the speakers? (As well as your ear and brain can do.)
    - Model the hidden dependence as
      - a linear combination of independent sources + noise
  - Many other similar examples in the areas of wireless, communications, signal restoration, etc.



Examples (cont'd)

- The Perception/AI viewpoint:
  - It is a complex world; one cannot model everything in detail
  - Rely on ***probabilistic models*** that explicitly account for the variability
  - Use the laws of probability to make inferences. E.g.,
    - P(burglar | alarm, no earthquake) is high
    - P(burglar | alarm, earthquake) is low
  - There is a whole field that studies "perception as Bayesian inference"
  - In a sense, perception really is "confirming what you already know."
  - priors + observations = robust inference



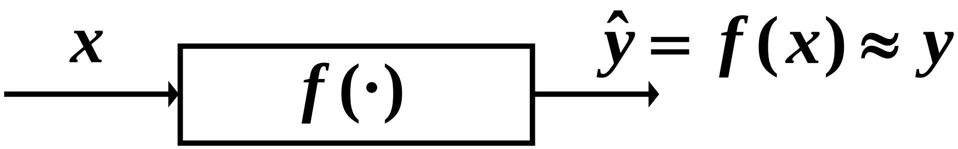
Examples (cont'd)

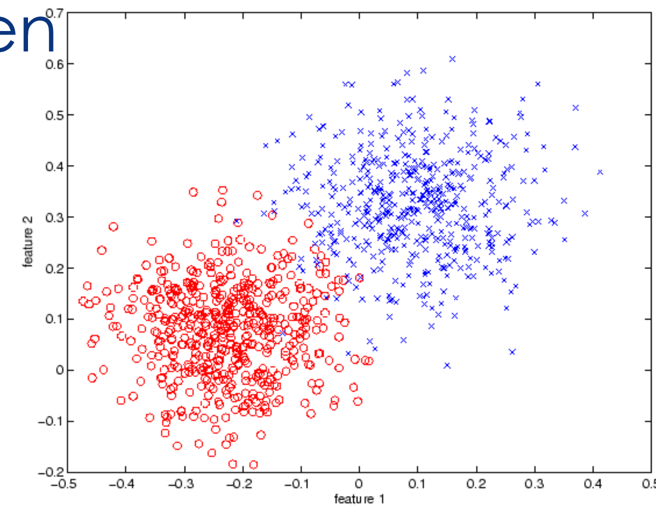
- The Communications Engineering viewpoint:
  - Detection problems:

  - You observe Y and know something about the statistics of the channel. What was X?
  - This is the canonical *detection problem.*
  - For example, face detection in computer vision: "I see pixel array Y. Is it a face?"

What is Statistical Learning?

- Goal: Given a relationship between a feature vector x and a vector y, and data samples (x_i, y_i), find an approximating function $f(x) \approx y$

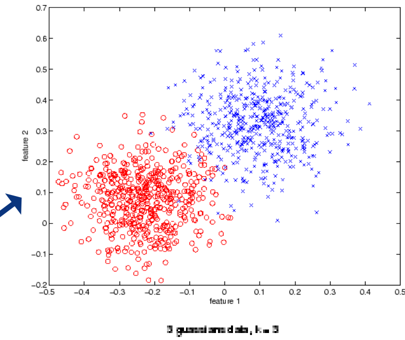


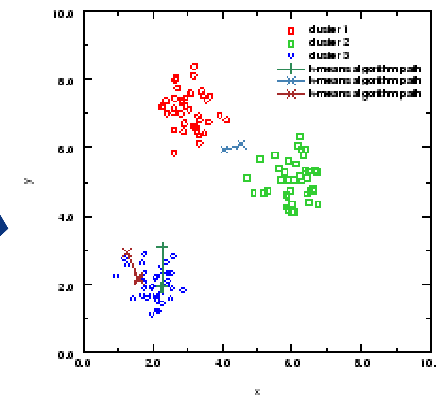


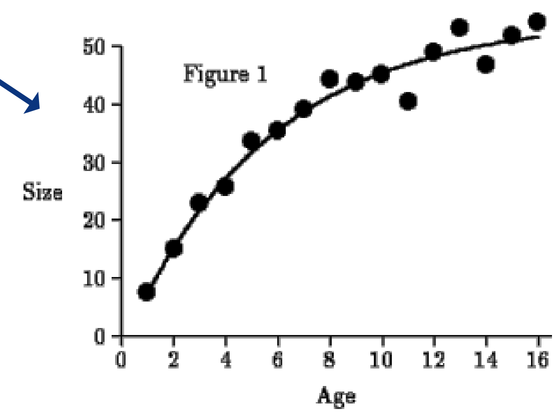
- This is called training or learning.
- Two major types of learning:
  - Unsupervised: only X is known, usually referred to as clustering.
  - Supervised : both $\boldsymbol{X}$ and target value $\boldsymbol{Y}$ are known during training, only $\boldsymbol{X}$ is known at test time. Usually referred to as classification or regression.

Supervised Learning

- X can be anything, but the type of known data Y dictates the type of supervised learning problem
  - Y in {0,1} is referred to as Detection or Binary Classification
  - Y in {0, ..., M-1} is referred to as (M-ary) Classification
  - Y continuous is referred to as Regression
- Theories are quite similar, and algorithms similar most of the time
- We will usually emphasize classification, but will talk about regression when particularly insightful





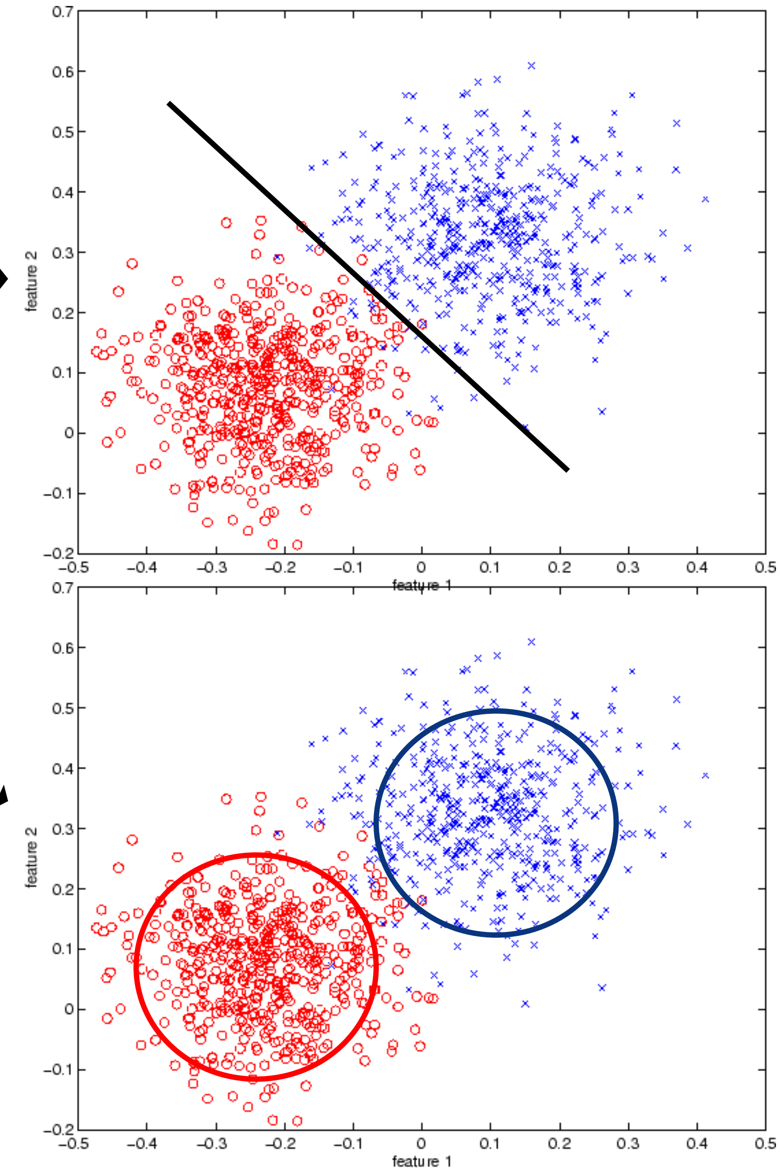


Example

- Classification of Fish:
  - Fish roll down a conveyer belt
  - Camera takes a picture
  - Decide if is this a salmon or a sea-bass?
- Q: What is X? E.g. what features do I use to distinguish between the two fish?
- This is somewhat of an art-form. Frequently, the best is to ask domain experts.
- E.g., expert says use *overall length* and *width of scales*.

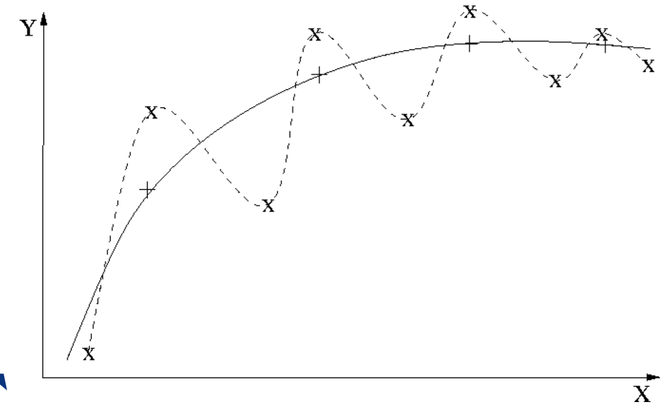
Classification/Detection

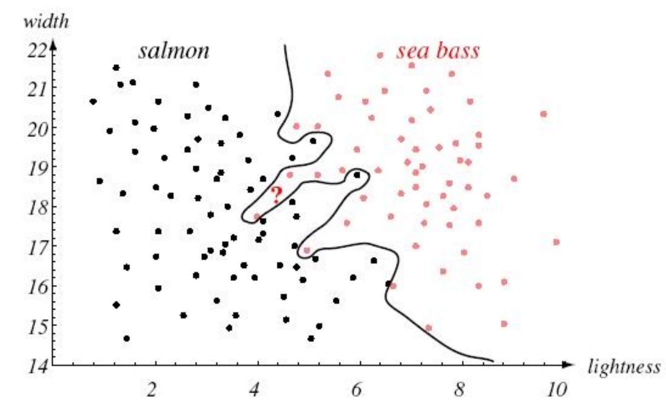
- Two major types of classifiers
- Discriminant: determine the decision boundary in feature space that best separates the classes;
- Generative: fit a probability model to each class and then compare the probabilities to find a decision rule.
- A lot more on the relationship between these two approaches later!



Caution

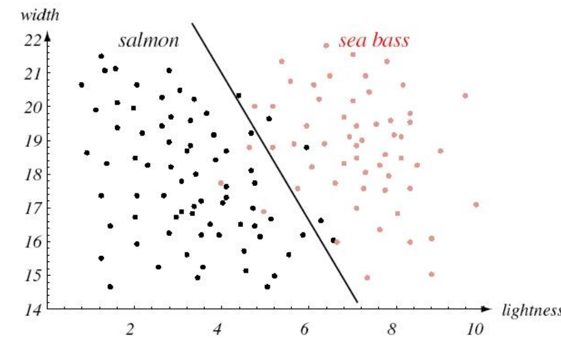
- How do we know learning has worked?
- We care about generalization, i.e. accuracy outside the training set
- Models that are "too powerful" can lead to over-fitting:
  - E.g. in regression one can always exactly fit n pts with polynomial of order n-1.
  - Is this good? how likely is the error to be small outside the training set?
  - Similar problem for classification
- Fundamental Rule: only hold-out test-set performance results matter!!!

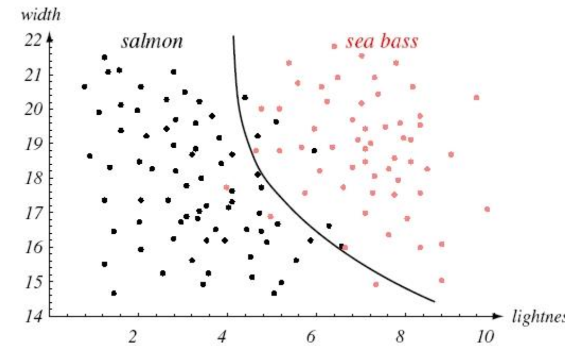


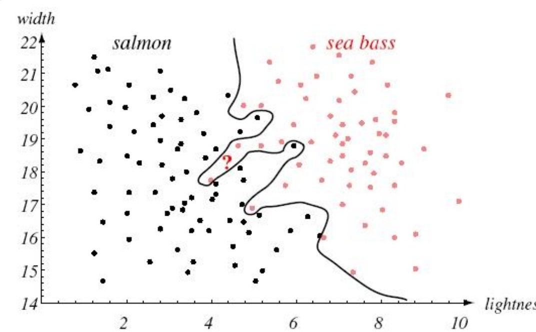


Generalization

- Good generalization requires controlling the trade-off between training and test error
  - training error large, test error large
  - training error smaller, test error smaller
  - training error smallest, test error largest
- This trade-off is known by many names
- In the generative classification world it is usually due to the bias-variance trade-off of the class models







Any questions?