Machine Learning I: Supervised Methods

B. Keith Jenkins

Announcements

- My office hours are:
  - M W 11:00 AM - 12:30 PM
    - EEB 403 and over zoom
- TA office hours will be announced
- Discussion session ~~tomorrow~~ Friday
  - 11:00 - 11:50 AM, OHE 122
- Homework 1 will be posted Friday 1/19

Reading

- Bishop
  - 4.1.0 – 4.1.2 (Discriminant functions)
  - Appendix C (Properties of Matrices) - read any sections of material you haven't already had

Today's Lecture

- Syllabus: administrative parts (finish)
- Basic concepts in ML
  - Classification and regression
- Key elements of ML systems
- Syllabus: course outline

Assessment Tool (assignments)	% of Grade
Homework	20
Course project	24
Midterm exam	24
Final exam	24
Class participation	8
TOTAL	100

Extra credit homework problems (throughout the semester) will be tallied separately from the above scores, and will determine students' grades for borderline cases (*i.e.*, students near a border between two letter grades).

Assignment Submission Policy

Submit your homework by uploading one pdf file of your solution, and one computer-readable pdf file of all your code (for assignments with computer problems), to the Assignment Dropbox on the D2L website. A computer-readable pdf can be generated by converting an editable text document to pdf; Jupyter notebooks can be converted to pdf (or first html) using nbconvert. Scanned documents and screen-shots are not computer readable.

Due date and time will be stated on each homework assignment. The late submission policy will be posted on the D2L web site, and will allow for a given number of free late days. Late penalties will be uniformly applied to everyone. Exceptions will only be granted for unusual or emergency situations (e.g., documented emergencies or medical conditions).

Grading Timeline

Graded assignments including comments will be available on the D2L website as soon as grading is completed and verified (typically ~2 weeks after the due date).

Additional Policies

Policy on Collaboration and Individual Work in this Class

Collaboration on techniques for solving homework assignments and computer problems is allowed, and can be helpful; however, each student is expected to work out, code, and write up his or her own solution. Use of other solutions to homework assignments or computer problems, from any source including other students, before the assignment is turned in, is not permitted.

For class projects, general collaboration to resolve issues, or to clarify technical material, is allowed. Use of internet as well as journal and conference literature is encouraged. However, each student (or team) does their own work and writes up their own report. The author(s) of the report are presenting themselves as having done the work described in the report. Any reported work, explanations, information, or code that is obtained from others must be cited as such; instructions for doing this will be given with the project assignment. Including such work in the report or code without citing it amounts to plagiarism.

Please also see below for additional policies that apply to all USC classes.

Statement on Academic Conduct and Support Systems

Academic Conduct:

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in SCampus in Part B, Section 11, "Behavior Violating University Standards" policy.usc.edu/scampus-part-b. Other forms of academic dishonesty are equally unacceptable. See additional information in SCampus and university policies on scientific misconduct, policy.usc.edu/scientific-misconduct.

Support Systems:

Counseling and Mental Health - (213) 740-9355 – 24/7 on call
studenthealth.usc.edu/counseling
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

National Suicide Prevention Lifeline - 1 (800) 273-8255 – 24/7 on call
suicidepreventionlifeline.org
Free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week.

Relationship and Sexual Violence Prevention Services (RSVP) - (213) 740-9355(WELL), press "0" after hours – 24/7 on call
studenthealth.usc.edu/sexual-assault
Free and confidential therapy services, workshops, and training for situations related to gender-based harm.

Office of Equity and Diversity (OED) - (213) 740-5086 | Title IX – (213) 821-8298
equity.usc.edu, titleix.usc.edu
Information about how to get help or help someone affected by harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants.

Reporting Incidents of Bias or Harassment - (213) 740-5086 or (213) 821-8298
usc-advocate.symplicity.com/care_report
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office of Equity and Diversity |Title IX for appropriate investigation, supportive measures, and response.

The Office of Disability Services and Programs - (213) 740-0776
dsp.usc.edu
Support and accommodations for students with disabilities. Services include assistance in providing readers/notetakers/interpreters, special accommodations for test taking needs, assistance with architectural barriers, assistive technology, and support for individual needs.

USC Campus Support and Intervention - (213) 821-4710
campussupport.usc.edu
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity at USC - (213) 740-2101
diversity.usc.edu

Information on events, programs and training, the Provost's Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
dps.usc.edu, emergency.usc.edu
Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety - UPC: (213) 740-6000, HSC: (323) 442-120 – 24/7 on call
dps.usc.edu
Non-emergency assistance or information.

Basic Concepts in Supervised ML

Regression

Outputs are typically continuous: $y_i \in \mathbb{R}$ (given outputs)

$\hat{y}_i(\underline{x}_i) = \hat{f}(\underline{x}_i) \in \mathbb{R}$ (output predictions)

Ex Predict housing prices

Data

Define inputs and outputs (possibly tentative)

Gather data

e.g.: from a real estate firm

from internet

⋮

More generally: data from sensors in the physical world.

pre-processing / data engineering:

- Inspect / clean the data
- Deal with any missing data
- Normalization (e.g., for very differing i/p values, or differing units on i/p's.)
- Categorical data (e.g., style of house: contemporary, colonial, etc.) to numeric data.

Exploratory data analysis (EDA)

Correlations between each input (e.g., living area) and house price

Correlation between inputs (e.g., living area and location)

Features

- Feature engineering
  - eliminate some features
  - create some new features from existing features or input values.
  - transformations of feature space. (e.g., PCA)

feature design

- How to come up with what the new features should be?
  - use prior knowledge about the problem (e.g., (liv. area) x (interior height) = interior volume.)
  - e.g., extracting useful features from audio streams for speech recognition.

dR

feature learning: - use an automated system (ML algorithm of sorts) to learn a good set of features

Now we have a set of features, e.g.:

x_1 = living area (sq. ft.)
x_2 = number of rooms
x_3 = longitude } (GPS data)
x_4 = latitude }
x_5 = year built
x_6 = park space
x_7 = # of parking spaces
x_8 = public school rating
x_9 = average salary of the region
⋮

$$\text{park space} \triangleq \frac{\text{area of nearest park}}{\text{distance to nearest park}}$$

Each house i in our dataset $\rightarrow$ feature vector $\underline{x}_i$, output value (\$) y_i.

Dataset: $\mathcal{D} = \{ \underline{x}_i, y_i \}_{i=1}^{N}$, N = # houses in our dataset.

Want: relationship $\underline{x} \longrightarrow \hat{y}$ = the output prediction for $\underline{x}$.

How come up with mapping $\underline{x} \to \hat{y}$?

Model

→ (i) Decide on a set of functions to use (hypothesis set)

→ (ii) Choose the function that gives best fit.

e.g.:

(i) Linear set of fcns. $\hat{f}(\underline{x}) = \sum_{j=0}^{D} w_j x_j$, (or $\underline{\underline{W}}\,\underline{x}$)

$x_0 = 1$. ($\Rightarrow w_0$ is bias or offset term)

(iii) Need a loss function that defines "goodness" of fit

e.g.:

$$\ell(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2 \quad \text{(squared-error } (\ell_2) \text{ loss)}$$

Learning

⇒ Learn by optimization to find optimal $\hat{\underline{w}} = \begin{bmatrix} \hat{w}_0 \\ \hat{w}_1 \\ \vdots \\ \hat{w}_D \end{bmatrix}$

⇒ ML algorithm

Learning by optimization:

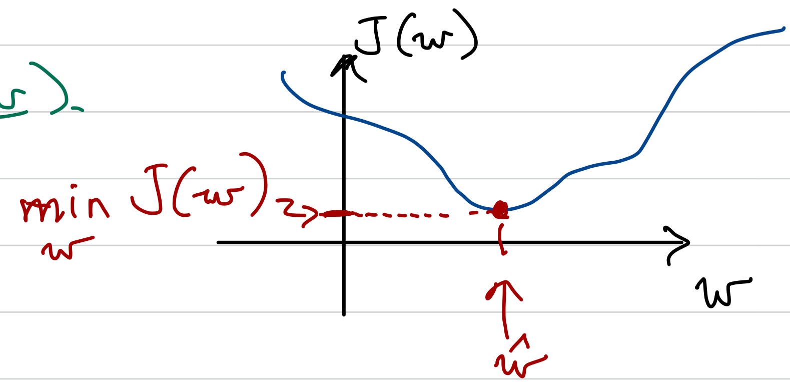
1. Need <u>criterion</u> or <u>objective</u> <u>function</u>, e.g.:

$$J(\underline{w}) = MSE = \frac{1}{N_{Tr}} \sum_{i=1}^{N_{Tr}} (\hat{y}_i - y_i)^2 , \quad \text{with } \hat{y}_i \triangleq \hat{f}(\underline{x}_i).$$

↑ sum over training data points in $\mathcal{D}_{Tr} \subset \mathcal{D}$

2. Find a minimum of $J(\underline{w})$ w.r.t. parameters $\underline{w}$

$$\Rightarrow \quad \hat{\underline{w}} = \underset{\underline{w}}{\operatorname{argmin}} \, J(\underline{w}).$$



$$\hat{\underline{w}} = \underset{\underline{w}}{\operatorname{argmin}} \, J(\underline{w})$$

Comment: $\mathcal{D}_{Tr}$ = training dataset

$\mathcal{D}_{Test}$ = test dataset

$$\mathcal{D}_{Tr} \cap \mathcal{D}_{Test} = \emptyset \qquad \text{(sets are disjoint)}$$

How to assess the system's prediction accuracy?

$\rightarrow$ Use $\mathcal{D}_{Test}$

i/p $\rightarrow \underline{x}_i \rightarrow \boxed{ML} \rightarrow \hat{f}(\underline{x}_i) = \hat{y}_i$

Compare $\hat{y}_i$ with y_i for all data pts. in $\mathcal{D}_{Test}$.

$\Rightarrow MSE_{Test}$

Post processing

Performance estimation

Performance measure

For classification problems:

Data
Features

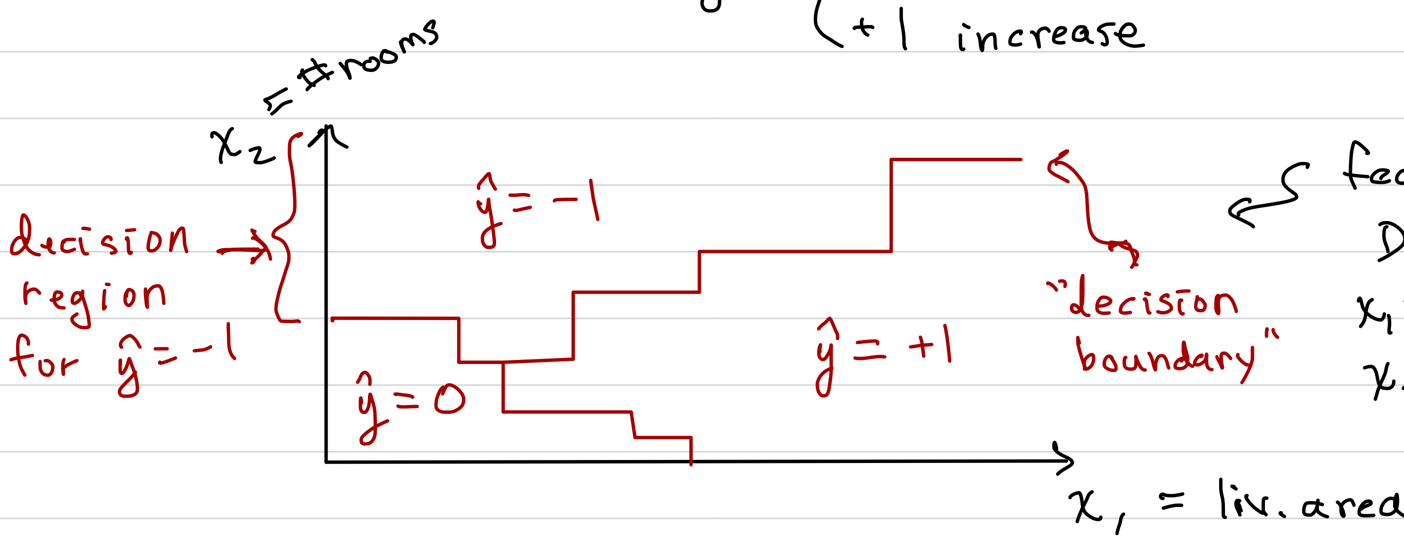
} similar to regression case

Mapping $\underline{x} \rightarrow \hat{y}$ is different

Ex: Predict change in housing price (now to future).

output class $\in$ {increase, decrease, constant} $\Rightarrow$ 3-class problem

encode as: $y = \begin{cases} -1 & \text{decrease} \\ 0 & \text{constant} \\ +1 & \text{increase} \end{cases}$



feature space
D=2 features
x_1 = living area
x_2 = # rooms

Key Elements of Machine Learning Systems - Summary

- Data - gathering, preprocessing, EDA (exploratory data analysis)
- Features - feature engineering, feature learning
- Models - hypothesis set (functions to try)
- Learning - optimization of $J(\underline{w})$.
- Prediction - output of ML system for any input $\underline{x}$.
- Post-processing - e.g., evaluating performance

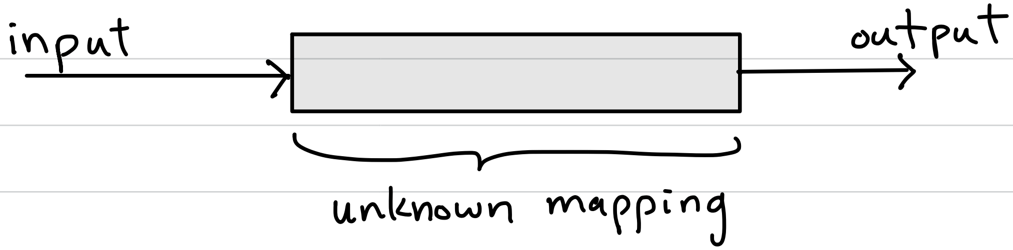
Included in these elements are

- Optimization
- Criterion functions and loss functions
- Performance measures and estimates

many choices for each

Supervised ML

Typical problem:



Use data — examples of (input, output) pairs — to estimate or model the unknown mapping, so that the system can <u>generalize</u>, that is, can estimate or predict outputs corresponding to previously unseen inputs.

Course outline (summary version)

> Order of topics covered will be slightly different than the order presented below
> Abbreviations: ML: machine learning; ANN: artificial neural networks

1. Course introduction
   - Basic concepts, ML paradigm, fundamental assumptions
2. Preliminaries
   - Discriminant functions; methods for mutlclass classification
   - Computational complexity
   - Convex functions
   - Fundamental assumptions
   - The learning problem
   - Criterion (objective) functions
   - Approaches to optimization
3. Learning and optimization 1
   - Gradient descent methods
   - Perceptron learning and convergence proof
   - Mean-squared-error algorithms for regression and classification
4. Nonlinear approaches for regression and classification
   - Nonlinear transformation by basis functions
   - Nonlinear transformation by kernel functions
5. Complexity in machine learning
   - Degrees of freedom and constraints
   - Regularization
   - Introduction to VC dimension
6. Learning and optimization 2
   - Optimization with constraints: Lagrangian techniques
   - Kernels and kernel methods
   - Support vector machines for classification
   - Support vector machines for regression
7. Validation and error estimation
   - Dataset usage: training, validation, test sets
   - Model selection and cross validation
8. Artificial neural networks 1
   - Single layer and multiple layer feedforward networks
   - Interpretations and capabilities
   - Learning algorithms
9. Artificial neural networks 2
   - ANNs as universal function approximators (proof by construction)

- Example: radial basis function networks
- Degrees of freedom and complexity in ANN

10. Feature selection and dimensionality reduction
    - Principal Components Analysis
    - Other linear transformations for 2-class and multiclass problems

---

probabilistic view

11. Bayes decision theory for classification
    - Theoretically optimal probabililstic classification (minimum error and minimum risk criteria)
12. Density estimation techniques for classification
    - Mathematical approach and convergence
    - Kernel density estimation and k-nearest neighbors estimation
    - Discriminative and generative approaches
13. Density estimation techniques for regression
    - Theoretically optimal probabilistic MSE regression
    - k-nearest neighbors regression
    - Computational complexity and speed-up techniques (regression and classification)
14. Parametric estimation techniques (as time permits)
    - For classification
    - For regression
15. Conclusion and summary
    - Relation of topics to each other
    - ML key elements in the design phase, learning phase, and prediction phase

pdf's ~~for~~ that underlies the data:

$$p(\underline{x} \mid S_k) = p(\underline{x} \mid S_k = \hat{y})$$

ex: $p(x_1 \mid S_1) = p(\text{living area} \mid \text{price increase})$