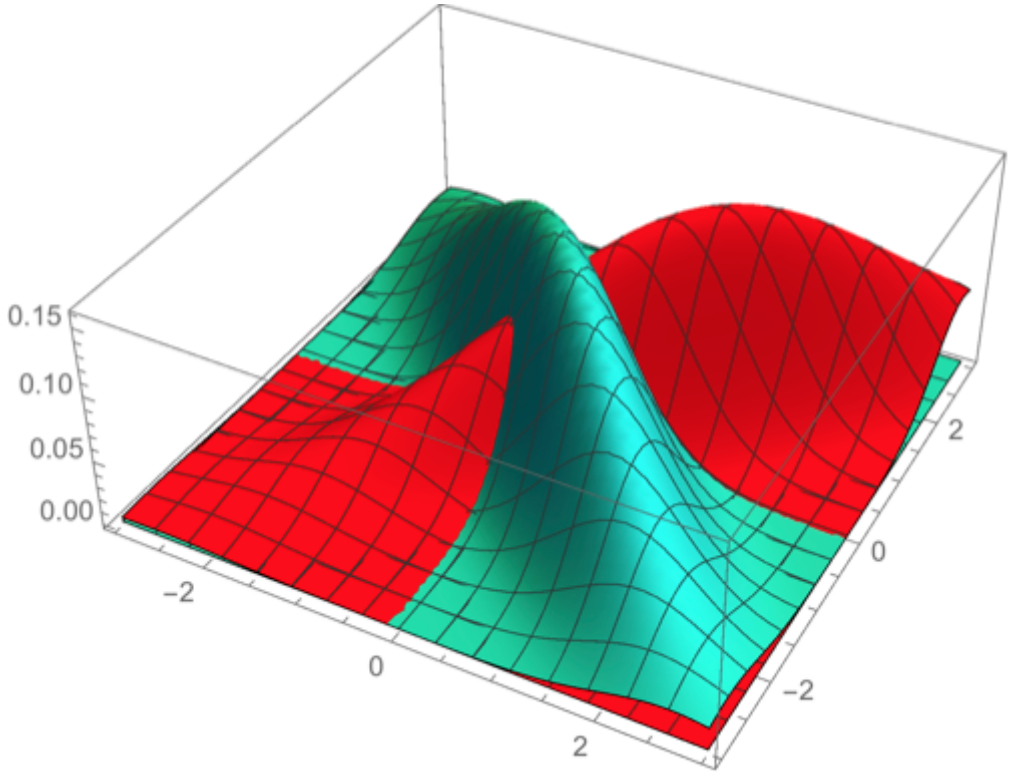


CS 189/289A Introduction to Machine Learning

Jonathan Shewchuk
jrs@cory.eecs.berkeley.edu
(Please send email only if you don't want the TAs to see it; otherwise, use Piazza.)

Spring 2017
Mondays and Wednesday, 6:30–8:00 pm
2050 Valley Life Sciences Building

This class introduces algorithms for *learning*, which constitute an important part of artificial intelligence.

Topics include

- classification: perceptrons, support vector machines (SVMs), Gaussian discriminant analysis (including linear discriminant analysis, LDA, and quadratic discriminant analysis, QDA), logistic regression, decision trees, neural networks, convolutional neural networks, nearest neighbor search;
- regression: least-squares linear regression, logistic regression, polynomial regression, ridge regression, Lasso;
- dimensionality reduction: principal components analysis (PCA), latent factor analysis; and
- clustering: *k*-means clustering, hierarchical clustering, spectral graph clustering.

Useful Links

- See the schedule of class and discussion section times and rooms.
- Access the CS 189/289A Piazza discussion group.
- If you want an instructional account, you can get one online. No more paper forms. Go to the same link if you forget your password or account name.

Prerequisites

- Math 53 (or another vector calculus course),
- Math 54 or 110 (or another linear algebra course),
- CS 70 (or another probability course).

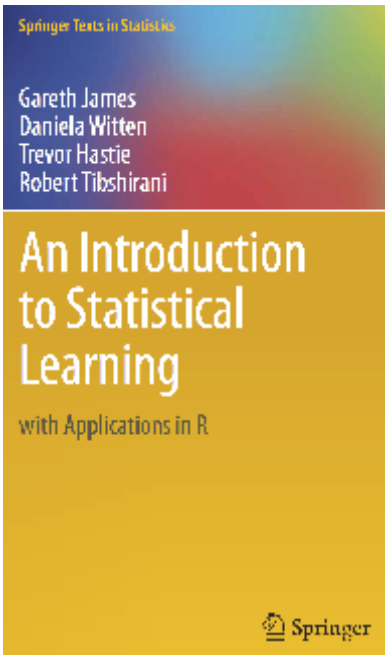
You should take these prerequisites quite seriously: if you don't have them, I strongly recommend not taking CS 189.

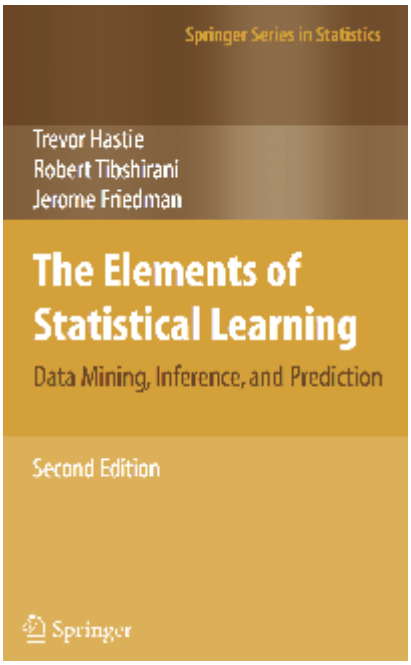
If you want to brush up on prerequisite material, Stanford's machine learning class provides nice reviews of linear algebra and probability theory. Here's a shorter summary of math for machine learning written by our TA Garrett Thomas. Other suggestions for review material appear in this Piazza post.

Textbooks

Both textbooks for this class are available free online. Hardcover and eTextbook versions are also available.

- Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, New York, 2013. ISBN # 978-1-4614-7137-0. See Amazon for hardcover or eTextbook.
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, second edition, Springer, 2008. See Amazon for hardcover or eTextbook.





Homework and Exams

Homework assignments will appear here.

You have a **total** of **5** slip days that you can apply to your semester's homework. We will simply not award points for any late homework you submit that would bring your total slip days over five.

The **Midterm** takes place on **Wednesday, March 15, in class**. Previous midterms are available: Spring 2013, Spring 2014, Spring 2015, Fall 2015, Spring 2016.

The **Final Exam** takes place on **Monday, May 8, 3–6 PM,** in a location to be determined later in the semester. CS 189 is in exam group 3. Previous final exams are available. Without solutions: Spring 2013, Spring 2014, Spring 2015, Fall 2015, Spring 2016. With solutions: Spring 2013, Spring 2014, Spring 2015, Fall 2015, Spring 2016.

Lectures

This list may change as the semester progresses.

Lecture 1 (January 18): Introduction. Classification, training, and testing. Validation and overfitting. Read ESL, Chapter 1. My lecture notes (PDF).

Lecture 2 (January 23): Linear classifiers. Predictor functions and decision boundaries. The centroid method. Perceptrons. Read parts of the Wikipedia Perceptron page. Optional: Read ESL, Section 4.5–4.5.1.

Lecture 3 (January 25): Gradient descent, stochastic gradient descent, and the perceptron learning algorithm. Feature space versus weight space. The maximum margin classifier, aka hard-margin support vector machine (SVM). Read ISL, Section 9–9.1.

Lecture 4 (January 30): The support vector classifier, aka soft-margin support vector machine (SVM). Features and nonlinear decision boundaries. Read ESL, Section 12.2 up to and including the first paragraph of 12.2.1.

Lecture 5 (February 1): Machine learning abstractions: application/data, model, optimization problem, optimization algorithm. Common types of optimization problems: unconstrained, constrained (with equality constraints), linear programs, quadratic programs, convex programs. Optional: Read (selectively) the Wikipedia page on mathematical optimization.

Lecture 6 (February 6): Decision theory: the Bayes decision rule and optimal risk. Generative and discriminative models. Read ISL, Section 4.4.1.

Lecture 7 (February 8): Gaussian discriminant analysis, including quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA). Maximum likelihood estimation (MLE) of the parameters of a statistical model. Fitting an isotropic Gaussian distribution to sample points. Read ISL, Section 4.4. Optional: Read (selectively) the Wikipedia page on maximum likelihood.

Lecture 8 (February 13): Eigenvectors, eigenvalues, and the eigendecomposition. The Spectral Theorem for symmetric real matrices. The quadratic form and ellipsoidal isosurfaces as an intuitive way of understanding symmetric matrices. Application to anisotropic normal distributions (aka Gaussians). Read Chuong Do's notes on the multivariate Gaussian distribution.

Lecture 9 (February 15): Anisotropic normal distributions (aka Gaussians). MLE, QDA, and LDA revisited for anisotropic Gaussians. Read ISL, Sections 4.4 and 4.5.

February 20 is Presidents' Day.

Lecture 10 (February 22): Regression: fitting curves to data. The 3-choice menu of regression function + loss function + cost function. Least-squares linear regression as quadratic minimization and as orthogonal projection onto the column space. The design matrix, the normal equations, the pseudoinverse, and the hat matrix (projection matrix). Logistic regression; how to compute it with gradient ascent or stochastic gradient descent. Read ISL, Sections 4–4.3.

Lecture 11 (February 27): Newton's method and its application to logistic regression. LDA vs. logistic regression: advantages and disadvantages. ROC curves. Weighted least-squares regression. Least-squares polynomial regression. Read ISL, Sections 4.4.3, 7.1, 9.3.3; ESL, Section 4.4.1.

Lecture 12 (March 1): Statistical justifications for regression. The empirical distribution and empirical risk. How the principle of maximum likelihood motivates the cost functions for least-squares linear regression and logistic regression. The bias-variance decomposition; its relationship to underfitting and overfitting; its application to least-squares linear regression. Read ESL, Sections 2.5 and 2.9. Optional: Read the Wikipedia page on the bias-variance trade-off.

Lecture 13 (March 6): Ridge regression: penalized least-squares regression for reduced overfitting. How the principle of maximum *a posteriori* (MAP) motivates the penalty term (aka Tikhonov regularization). Kernels. Kernel ridge regression. The polynomial kernel. Read ISL, Sections 6.2–6.2.1 and ESL, Sections 12.3–12.3.2. Optional: This CrossValidated page on ridge regression is pretty interesting.

Lecture 14 (March 8): Kernel perceptrons. Kernel logistic regression. The Gaussian kernel. Subset selection. Lasso: penalized least-squares regression for reduced overfitting and subset selection. Read ISL, Sections 6–6.1.2 and the last part of 6.1.3 on validation; and ESL, Sections 3.4–3.4.3. Optional: Read ISL, Section 9.3.2 if you're curious about kernel SVM.

Lecture 15 (March 13): Decision trees; algorithms for building them. Entropy and information gain. Read ISL, Sections 8–8.1.

The Midterm takes place in class on **Wednesday, March 15**. You are permitted one "cheat sheet" of letter-sized (8½" × 11") paper.

Lecture 16 (March 20): More decision trees: multivariate splits; decision tree regression; stopping early; pruning. Ensemble learning: bagging (bootstrap aggregating), random forests. Read ISL, Section 8.2.

Lecture 17 (March 22): Neural networks. Gradient descent and the backpropagation algorithm. Read ESL, Sections 11.3–11.4. Optional: I've heard positive recommendations for Welch Labs' video tutorial Neural Networks Demystified on YouTube. Also of special interest is this Javascript neural net demo that runs in your browser.

March 27–31 is Spring Recess.

Lecture 18 (April 3): Neuron biology: axons, dendrites, synapses, action potentials. Differences between traditional computational models and neuronal computational models. Backpropagation with softmax outputs and logistic loss. Unit saturation, aka the vanishing gradient problem, and ways to mitigate it. Heuristics for avoiding bad local minima.

Lecture 19 (April 5): Heuristics for avoiding bad local minima. Heuristics for faster training. Heuristics to avoid overfitting. Convolutional neural networks. Neurology of retinal ganglion cells in the eye and simple and complex cells in the V1 visual cortex. Read ESL, Sections 11.5 and 11.7. Optional: A fine paper on heuristics for better neural network learning is Yann LeCun, Leon Bottou, Genevieve B. Orr, and Klaus-Robert Müller, "Efficient BackProp," in G. Orr and K.-R. Müller (Eds.), *Neural Networks: Tricks of the Trade*, Springer, 1998. Also of special interest is this Javascript convolutional neural net demo that runs in your browser.

Lecture 20 (April 10): Unsupervised learning. Principal components analysis (PCA). Derivations from maximum likelihood estimation, maximizing the variance, and minimizing the sum of squared projection errors. Eigenfaces for face recognition. Read ISL, Sections 10–10.2 and the Wikipedia page on Eigenface.

Lecture 21 (April 12): The singular value decomposition (SVD) and its application to PCA. Clustering: *k*-means clustering aka Lloyd's algorithm; *k*-medoids clustering; hierarchical clustering; greedy agglomerative clustering. Dendrograms. Read ISL, Section 10.3.

Lecture 22 (April 17): Spectral graph partitioning and graph clustering. Relaxing a discrete optimization problem to a continuous one. The Fiedler vector, the sweep cut, and Cheeger's inequality. The vibration analogy. Greedy divisive clustering. The normalized cut and image segmentation. Read my survey of *Spectral and Isoperimetric Graph Partitioning*, Sections 1.2–1.4, 2.1, 2.2, 2.4, 2.5, and optionally A and E.2. For reference: Jianbo Shi and Jitendra Malik, *Normalized Cuts and Image Segmentation*, IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(8):888–905, 2000.

Lecture 23 (April 19): Graph clustering with multiple eigenvectors. Latent factor analysis (aka latent semantic indexing). Nearest neighbor classification and its relationship to the Bayes risk. The geometry of high-dimensional spaces. The exhaustive algorithm for *k*-nearest neighbor queries. Read ISL, Section 2.2.3. Optional: Read the Wikipedia page on latent semantic analysis. For reference: Andrew Y. Ng, Michael I. Jordan, Yair Weiss, *On Spectral Clustering: Analysis and an Algorithm*, Advances in Neural Information Processing Systems 14 (Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors), pages 849–856, the MIT Press, September 2002.

Lecture 24 (April 24): Speeding up nearest neighbor queries. Voronoi diagrams, order-*k* Voronoi diagrams, and point location. *k*-d trees. Application of nearest neighbor search to the problem of *geolocalization*: given a query photograph, determine where in the world it was taken. For reference: the IM2GPS web page, which includes links to the paper.

Lecture 25 (April 26): Guest lecture?

The Final Exam takes place on **Monday, May 8, 3–6 PM** in a location to be determined later in the semester. (CS 189 is in exam group 3.)

Discussion Sections and Teaching Assistants

Sections begin to meet on January 24. A schedule will appear here.

Grading

- **40%** for homeworks.
- **20%** for the Midterm.
- CS 189: **40%** for the Final Exam.
- CS 289A: **20%** for the Final Exam.
- CS 289A: **20%** for a project.

Supported in part by the National Science Foundation under Awards CCF-0430065, CCF-0635381, IIS-0915462, and CCF-1423560, in part by a gift from the Okawa Foundation, and in part by an Alfred P. Sloan Research Fellowship.