CS 189: Introduction to Machine Learning

Scribe: Tyler Nguyen

Lecture 1: January 18, 2017

Administrative Matters

Website: `http://www.cs.berkeley.edu/~jrs/189`
Questions: Please use Piazza, not email.
For personal matters only, `jrs@cory.eecs.berkeley.edu`.

Prerequisites

- Math 53 (vector calculus)
- Math 54 or 110 (linear algebra)
- CS 70 (probability)
- Not CS 188

Grading

- 40% 7 Homeworks: Late policy: 5 slip days total.
- 20% Midterm: Wednesday, March 15, in class.
- 40% Final Exam: Moved to Monday, May 8, 3-6 PM (Exam group 3).

Cheating

- Discussion of HW problems is encouraged.
- All homeworks, especially programming, must be written individually.
- We will actively check for plagiarism.
- Typical penalty is a large negative score, but I reserve the right to give an instant F for even one violation, and always give an F for two.

1 Core Material

- Finding patterns in data; using them to make predictions.
- Models and statistics help us understand patterns.
- Optimization algorithms "learn" the patterns.

2 Classifiers

- Decision boundary: Separating boundary between two classifications.
- Nearest neighbor: Classify according to nearest data point in training set. Possibly overfitting.
- Linear: Linear decision boundary.
- k-nearest neighbor cluster: Classify according to k nearest data point(s) in training set. "Smoother" decision boundary.

2.1 Classifiying digits

Images (arrays of pixel intensity values) are points in $m \times n$-dimensional space:

$$\begin{bmatrix} 3 & 3 & 3 & 3 \\ 0 & 0 & 2 & 3 \\ 0 & 0 & 1 & 3 \\ 3 & 3 & 3 & 3 \end{bmatrix} \rightarrow \begin{bmatrix} 3 \\ 3 \\ 3 \\ 3 \\ 0 \\ 0 \\ 2 \\ 3 \\ 0 \\ 0 \\ 1 \\ 3 \\ 3 \\ 3 \\ 3 \\ 3 \end{bmatrix}$$

Linear decision boundary is a hyperplane.

2.2 Validation

- **Train** a classifier: it **learns** to distinguish 7 from not 7.
- **Test** the classifier on new images.

2.2.1 2 kinds of error

1. **Training set error**: fraction of training images not classified correctly.
2. **Test set error**: fraction of misclassified new images, not used during training.

- **outliers**: points whose labels are atypical.
- **overfitting**: when the test error deteriorates because the classifier becomes too sensitive to outliers or other spurious patterns.

Most ML algorithms have a few **hyperparameters** that control over/underfitting, e.g. k in k-nearest neighbors. We select them by **validation**:

- Hold back a subset of training data, called the **validation set**.
- Train the classifier multiple times with different hyperparameter settings.
- Choose the settings that work best on validation set.

Now we have 3 sets:

- **training set** used to learn model weights.
- **validation set** used to tune hyperparameters, choose among different models.
- **test set** used as final evaluation of model. Kept in a vault. Run once, at the very end.

`Kaggle.com`:

- Runs ML competitions, including our HWs
- We use 2 data sets:
  - "public" set results available during competition
  - "private" set revealed after due date