Human Activity Recognition using Smartphone Sensors

Jessica Moore, Binghai Ling
jmoore5@stanford.edu, binghai@stanford.edu

Summary
- Cell phones are becoming ubiquitous and can reasonably be used to track daily activity
- We predicted the activity a person was taking, based on signals from their cell phone
- Using PCA and SVMs, we achieved a ~4% error rate with only about 200 features
- Most errors in the final models come from confusion of the 'sitting' and 'standing' classes

Data
- Source: UCI Machine Learning Repository
- Signals from waist-worn cellphones of 30 individuals performing 6 activities
- Features are summary statistics (e.g., mean, SD) of filtered accelerometer and gyroscope signals, obtained during a period of activity

Feature Engineering
Due to the size of the data set, variance seems likely to be the primary source of prediction error. To combat this, we implemented two unsupervised feature engineering methods:
- Principal Components Analysis
  Eigenvalue decomposition of:
  \[ \Sigma = \sum \sigma_i u_i^T u_i \]
  where \( \sigma_i \) is the jth PC
- Kernelized Principal Components Analysis
  Implicit eigenvalue decomposition of:
  \[ \Sigma = \sum \sigma_i k(x_i, x_j) \]
  Achieved via an eigenvalue decomposition of the kernel matrix

Results
- Features:
  - kPC-based features yielded high errors
  - PC-based features gave similar errors to the full, original data
  - SVMs performed well and used only ~200 PCs

Models
- K-Nearest Neighbors
  Classification based on k nearest training points
  \[ P(y^{(i)} = j | x = x^{(i)}) = \frac{1}{k} \sum_{j=1}^{k} P(y^{(i)} = j) \]
- Softmax Regression
  Classification based on maximum \( \theta^T x^{(i)} \)
  \[ P(y^{(i)} = j | x = x^{(i)}) = \frac{e^{\theta_j^T x^{(i)}}}{\sum_{j=1}^{C} e^{\theta_j^T x^{(i)}}} \]
  One vs. one
  One vs. all

Error Analysis & Future Work
- Low training error and distinctly higher test error suggests high variance
- Future work might include regularization or additional feature reduction to resolve this

Discussion
- Using PC-based features dramatically reduced the number of features, without affecting misclassification error.
- The success of SVMs and Softmax regression (with PC-based features) is likely because the classes are separate in the space of the first ~200-300 principal components.
- kPCA failure may indicate a lack of relationship between nonlinear variation and the response. This is surprising, given that prior authors had success with Gaussian-kernel SVMs.
- Primary errors in the final model arise from mislabeling of 'sitting' and 'standing' classes. These are often confused for one another.

Acknowledgements
We would like to thank Professor Ng and Professor Duch for their instruction throughout the quarter. We would also like to thank our Project TA, Francois, for his comments as we completed this project.

References: