Predicting Boston Rides

Gartheeban Ganeshapillai
CSAIL MIT
• Predicting the number of pickups in a 2 hr window within 250m radius from a location.

• Given
  – Pickups (P)
  – Dropoffs (D)
  – Events (E)
  – Twitter (TW)
  – Weather (W)
Nearby Pickups (N)
Nearby Pickups (N)
Nearby Pickups (N)
• Additional series created
  – Nearby pickups within 1.4*250m radius (N1)
  – Nearby pickups within 2*250m radius (N2)
  – Dropoffs elsewhere (D2)
  – Dropoffs at other specific locations, for e.g., at Logan international (D_1 D_2 \ldots D_{36})
Additional series created
– Nearby pickups within 1.4*250m radius (N1)
– Nearby pickups within 2*250m radius (N2)
– Dropoffs elsewhere (D2)
– Dropoffs at other specific locations, for e.g., at Logan international (D1 D2 … D36)
Method
• $y = f(x)$

• **Output:** $y$
  – Number of pickups

• **Inputs:** $x$
  – Pickups (P)
  – Dropoffs (D)
  – Twitter (TW)
  – Weather (W)
  – Nearby pickups (N1, N2)
  – Other dropoffs ($D_2$, $D_1$, $D_2$, … $D_{36}$)
• Cleaning, normalizing, centering data
• Centering
  – Day of the week, Hour of the day, and Month
• Extract time windows
• Torture the data until it speaks
Under-fitting
(too simple to explain the variance)

Appropriate-fitting

Over-fitting
(forcefitting -- too good to be true)
• Classical machine learning problems
  – Non-linear fit
  – Avoid over-fitting
• Bagged tree learner (random forest)
  – Builds a regression tree on a sample set, averages all the trees (hundreds) at the end
  – Powerful, non-linear, and avoids over-fitting.
• Time series - autocorrelation
• Another level of predictions on residuals ($\hat{y} - y$) in near-by hours
  – E.g., residuals the previous day and following day at the same hour, and residuals +/- 2 hour apart.
• 2 levels of corrections to handle autocorrelation on residuals.