Designing Information Devices and Systems II

Lecture 10A
k-means Clustering
k-means

Given: \( \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_m \in \mathbb{R}^n \)
Partition them into \( k << m \) groups

0) Guess cluster centers to initialize
1) Group points around nearest center
2) Update cluster centers by averaging within group
3) If centers have changed, repeat 1-3
k-means 1D example

\[ n = 1, m = 8, k = 2 \]
General k-means Algorithm

0) Initialize k cluster centers \( \vec{m}_1, \vec{m}_2, \ldots, \vec{m}_k \)

1) Assign points to cluster: point \( \vec{x} \) goes to cluster \( i \) if, 
\[
||\vec{x} - \vec{m}_i|| < ||\vec{x} - \vec{m}_j|| \quad \forall j \neq i
\]

2) Let \( S_i \) be the set of samples in cluster \( i \)
recompute cluster centers:
\[
\vec{m}_i = \frac{1}{|S_i|} \sum_{\vec{x} \in S_i} \vec{x}
\]

3) If any \( m_i \) has changed, repeat 1-3
Objective Function

Find the clustering of $\vec{x}_1, \cdots, \vec{x}_m$ into sets $S_1, \cdots, S_k$ which minimizes:

$$D = \sum_{i=1}^{k} \sum_{\vec{x} \in S_i} ||\vec{x} - \mu_i||$$

$$\mu_i = \frac{1}{|S_i|} \sum_{\vec{x} \in S_i} \vec{x}$$

While the algorithm decreases the objective, the objective is non-convex and can be stuck on local minima.

General problem is N-P Complete
Management of intersections with multi-modal high-resolution data

Ajith Muralidharan\textsuperscript{1}, Samuel Coogan\textsuperscript{2}, Christopher Flores, Pravin Varaiya \textsuperscript{*}

Sensys Networks, Inc, Berkeley, CA 94710, United States

1. Introduction

Urban traffic in the U.S. is regulated by 300,000 signalized intersections, whose performance is determined by their signal timing plans. The NCHRP report (Dowling and Ashiabor, 2012) found 98 percent of agencies use qualitative rather than data-driven methodologies for timing plans in the face of variability in demand and saturation rates. The report describes methodologies for robust operation of signal systems. However, current traffic signal control gives an overall grade of D+ (National Traffic Operations Coalition, 2007). The HR system in Danville, CA.  

Fig. 1. HR system in Danville, CA.
Traffic Patterns

What would k-means cluster to?

K = 2?

K = 4?
The procedure described above to design the splits for time interval $T_i$ uses only the average counts $C_l i$ in that interval. But the HR system provides real-time counts, which one could use to predict future volume. These predictions could be used to adapt the splits to take the predicted traffic into account. We use a principal component analysis (PCA) for prediction. To illustrate the idea consider the North–South through movement (see Fig. 4), averaged over 15 min, for an entire day for 164 weekdays during Dec 2014–May 2015. Each day $d$ gives a $96 = 4 \times 24$ dimensional vector, $v(d)$, from which we subtract the average $\overline{v} = \frac{1}{164} \sum_d v(d)$. We perform a PCA of the vectors $x(d) = \frac{v(d)}{\overline{v}}$. Fig. 8 displays the average, $\overline{v}$, and the top four principal components, as well as the PCA weights of each component, which rapidly decrease beyond four. Denote these components by $x_1, x_2, x_3, x_4$. Then the count vector $v(d)$ for any day $d$ is represented as

$$v(d) = \overline{v} + w_1(d) x_1 + w_2(d) x_2 + w_3(d) x_3 + w_4(d) x_4 + \epsilon(d),$$

where the $w_i(d)$ are the coefficients of the orthogonal projection of $\frac{v(d)}{\overline{v}}$ on the ‘basis’ vectors $x_1, x_2, x_3, x_4$ and $\epsilon(d)$ is the residual.

---

**Fig. 5.** Clustering of daily data for Dec 2014 to May 2015 in an intersection in Beaufort, SC.
3.1. Wasted green and max waiting time

Fig. 9 (left) is the histogram of wasted green in phase 4 within a cycle, defined as the time during which this phase is actuated and there is no vehicle present, but there is a vehicle waiting for the actuation of another phase. The histogram on the right is the maximum time a vehicle is waiting at the stop bar for the phase to be actuated. The histograms show that the median wasted green time per cycle is 6 s, and the median max wait time is 20 s; about 10% of cycles inflict a very large wait time, which may cause driver complaints.

3.2. Purdue coordination diagram (PCD)

The HR time series can be used to measure the quality of traffic progression at an intersection, and subsequently for signal offset re-timing calculations in order to coordinate signals at multiple traffic intersections to ensure smooth progression. Fig. 10 is the Purdue Coordination Diagram (PCD) (Bullock et al., 2008). It is a visual representation of the quality of traffic progression through the Danville intersection. It displays the histogram and the cumulative distribution function (CDF) of vehicle arrivals at the intersection relative to the beginning of the green signal for a chosen time interval (in this case, the morning peak, 7 am–10 am). The plots immediately reveal the percent arrivals on green and vehicle waiting time – two important metrics characterizing traffic progression. The PCD shows that 60 percent of vehicles from the East and 40 percent from the West arrive on red and have to wait at the intersection.

4. Safety

We describe applications of HR data from the Danville intersection to assess moving violations, near crashes, and pedestrian–vehicle interaction.

4.1. Moving violations

Measurements of a vehicle’s movement, its speed, and the signal phase, allows one to determine if its driver is making a moving violation. Three violations compromise safety: the vehicle is speeding, running a red light, or making a right turn on red (RTOR) without stopping.

Fig. 11 plots vehicles that leave the stop bar around the time the signal transitions through yellow before turning red. (The yellow signal in the Danville intersection lasts 3 s.) Vehicle speeds are plotted against the time of intersection entry, relative to the yellow signal start time. The speed measurements are quantized because of the detectors finite sampling frequency: the estimates are accurate to within ±3 mph for speeds under 20 mph but higher speeds have larger quantization errors. The red circles in Fig. 11 (top) capture five vehicles driven dangerously across the intersection at high speed within 3 s after the signal turns red—a rate of one violation per hour in the east direction. The high speed suggests that drivers find themselves in...