Balloon Ranking Notes

Pushpendre Rastogi

Jeremy Silver

Johns Hopkins HLTCOE

January 20, 2017

1 Definitions and Computation

Suppose we have some points in $\mathbb{R}^d$, some of which are "seeds" known to have some property of interest, and we wish to rank some other points based on their likelihood of having this property. A reasonable assumption to make is that points that are close to seeds should be ranked higher than points that are farther away.

Let $\mathcal{P} \subseteq \mathbb{R}^d$ be the set of all points to be ranked, and let P be its cardinality. Let $\mathcal{S} \subseteq \mathbb{R}^d$ be the set of all seeds, and let S be its cardinality. Using i to index the points and j to index the seeds, let d_{ij} be the distance of point p_i from seed s_j.

Let $\tilde{r}$ be the length-K sequence of unique d_{ij} values. We assume that $\tilde{r}$ is sorted in ascending order, i.e. $\tilde{r}_1 < \tilde{r}_2 < \ldots < \tilde{r}_{\mathrm{K}}$.[1] Define c_{ik} to be the number of seeds s_j such that $d_{ij} \leq r_k$. Let $D \in \mathbb{R}^{\mathrm{P}\times\mathrm{S}}$ and $C \in \mathbb{N}^{\mathrm{P}\times\mathrm{K}}$ be matrices containing d_{ij} and c_{ik} respectively.

Definition 1. Let $\tilde{r}_0 = 0 < \tilde{r}_1 < \tilde{r}_2 < \ldots < \tilde{r}_K < \tilde{r}_{K+1} = \infty$. Then we define the **inflation ranking** $>$ by:

$$p_i > p_{i'} \quad \text{iff} \quad \exists k \text{ such that } c_{ik} > c_{i'k} \text{ and } \forall k' < k,\, c_{ik'} = c_{i'k'}.$$

Theorem 1. The inflation ranking is a strict weak ordering.

Proof.

- **Irreflexivity.** $p_i > p_i$ implies $c_{ik} > c_{ik}$ for some k, which is impossible.
- **Asymmetry.** $p_i > p_{i'}$ implies $c_{ik} > c_{i'k}$ for some k, while (without loss of generality) $p_i < p_{i'}$ implies $c_{ik'} \leq c_{i'k'}$ for all $k' \leq k$. These two statements are contradictory.
- **Transitivity**. Suppose $p_\ell > p_m$ and $p_m > p_n$. Then

  1) $\exists k_1$ s.t. $c_{\ell k_1} > c_{mk_1}$ and $\forall k < k_1$, $c_{\ell k} = c_{mk}$,

  2) $\exists k_2$ s.t. $c_{mk_2} > c_{nk_2}$ and $\forall k < k_2$, $c_{mk} = c_{nk}$.

[1] $\tilde{r}$ does not contain duplicates; thus it can be sorted into a strictly increasing sequence.

If $k_2 > k_1$, then $c_{\ell k_1} > c_{m k_1}$ by (1), and $c_{m k_1} = c_{n k_1}$ by (2), so $c_{\ell k_1} > c_{n k_1}$. And for all $k < k_1$, we have $c_{\ell k_1} = c_{m k_1}$ by (1), and $c_{m k_1} = c_{n k_1}$ by (2), so the result holds.

If $k_2 = k_1$, the result follows immediately from the conjunction of both statements in (1) and (2), along with transitivity of inequality.

If $k_2 < k_1$, then $c_{\ell k_2} = c_{m k_2}$ by (1), and $c_{m k_2} > c_{n k_2}$ by (2), so $c_{\ell k_2} > c_{n k_2}$. And for all $k < k_2$, we have $c_{\ell k_2} = c_{m k_2}$ by (1), and $c_{m k_2} = c_{n k_2}$ by (2), so the result holds.

- **Transitivity of incomparability.** Suppose neither $p_\ell > p_m$ nor $p_\ell < p_m$, and neither $p_m > p_n$ nor $p_m < p_n$ hold. It follows that for all k, $c_{\ell k} = c_{mk} = c_{nk}$, and the transitivity of equality yields incomparability of p_ℓ and p_n.

□

Definition 2. Let $\tilde{r}_0 = 0 < \tilde{r}_1 < \tilde{r}_2 < \ldots < \tilde{r}_K < \tilde{r}_{K+1} = \infty$. Then we define the **deflation ranking** $>$ by:

$$p_j > p_{j'} \quad \text{iff} \quad \exists k \text{ such that } c_{jk} > c_{j'k} \text{ and } \forall k' > k,\, c_{jk'} = c_{j'k'}.$$

Theorem 2. The deflation ranking is a strict ordering.

Proof. Analogous to Theorem 1. □

Collectively we assign to these inflation and deflation rankings the term "balloon ranking."

1.1 Computation of Balloon Ranking

Naive Method: To compute a balloon ranking, it suffices to compute the matrix C and then apply the definition of the ranking directly. The following is a naïve method for computing C:

```
for k, r_k in enumerate(r_tilde):
    C[:, k] = (D > r_k).sum(axis=1)
```

That is, for each unique radius $\tilde{r}_k$, we count the total number of seeds within the $\tilde{r}_k$-ball of each point. The overall complexity of this method is $O(\mathrm{KSP})$ operations. Since K can be equal to SP, this method is impractical.

Fast Computation: Let m_{ij} be the index such that $\tilde{r}_{m_{ij}} < d_{ij} \leq \tilde{r}_{m_{ij}+1}$. By construction, seed j will contribute a value of 1 to c_{ik} for all $\tilde{r}_k \geq \tilde{r}_{m_{ij}}$, i.e. $c_{ik} = \sum_{j \in \{1,\ldots,\mathrm{S}\}} \mathbb{I}[m_{ij} \leq k]$. Clearly, when $k = \mathrm{K}$ then $c_{ik} = \mathrm{S}$.

Define $M \in \mathbb{N}^{\mathrm{P}\times\mathrm{S}}$ to be the matrix that contains m_{ij}. This matrix can be computed in time $O(\mathrm{SP}\log(\mathrm{K}))$, which is efficient even if $\mathrm{K} = \mathrm{SP}$. After computing M, we can sort each row of M in the ascending order using $O(\mathrm{PS}\log(\mathrm{S}))$ operations to create the matrix $\tilde{M}$. $\tilde{M}$ is a compressed version of C. For example, if $k < \tilde{m}_{i1}$ then c_{ik} equals 0, and if $k \geq \tilde{m}_{i\mathrm{S}}$ then c_{ik}

equals S. Finally, if $\tilde{m}_{ij} \leq k < \tilde{m}_{i(j+1)}$ then $c_{ik} = j$; therefore c_{ik} can be computed using a binary search on the rows of $\tilde{M}$.

The inflation ranking uses the following decision function: $p_i > p_{i'}$ if there exists k such that $c_{ik} > c_{i'k}$ and $c_{ik'} = c_{i'k'}$ $\forall k' < k$. Assuming access to $\tilde{M}$, this decision function can be computed efficiently by comparing the prefixes of the sequences $\tilde{m}_{i:}$ and $\tilde{m}_{i':}$ and by finding the first prefix where the two sequences differ. Let j be the first index such that $\tilde{m}_{ij} \neq \tilde{m}_{i'j}$. If $\tilde{m}_{ij} > \tilde{m}_{i'j}$, then $p_i < p_{i'}$; otherwise $p_i > p_{i'}$.

1.2 Binary Labeled Seeds

We now consider the case where our seed points contain binary labels. That is, some seeds are "positive" examples and others are "negative" examples, and we wish to consider both nearness to positive seeds and distance from negative seeds when computing our ranking. The definitions of inflation and deflation only make use of the distances, d_{ij}, and the scores at each distance, c_{ik}; they are agnostic to how the c_{ik} are computed, so we are free to modify this in order to accommodate binary labels.

Let $\mathcal{J}^+$ and $\mathcal{J}^-$ be the partition of $\{1, 2 \ldots, \mathrm{S}\}$ into sets of indices corresponding to the positive and negative seeds, respectively. Let $\lambda \in [0, 1]$ be a weight parameter that determines the relative influence of positive and negative seeds. Let $\tilde{r}$ be just as before the ascending sequence of all seed-point distances. Then the weighted score matrix C can be computed via

$$c_{ik} = \lambda \sum_{j \in \mathcal{J}^+} I[m_{ij} \leq k] - (1 - \lambda) \sum_{j \in \mathcal{J}^-} I[m_{ij} \leq k].$$

We see that the case $\lambda = 1$ is identical to the earlier version of the score that only considers positive seeds, while $\lambda = 0$ only considers negative seeds, and $\lambda = 0.5$ assigns equal weight to both.

TODO: computation of weighted ranking

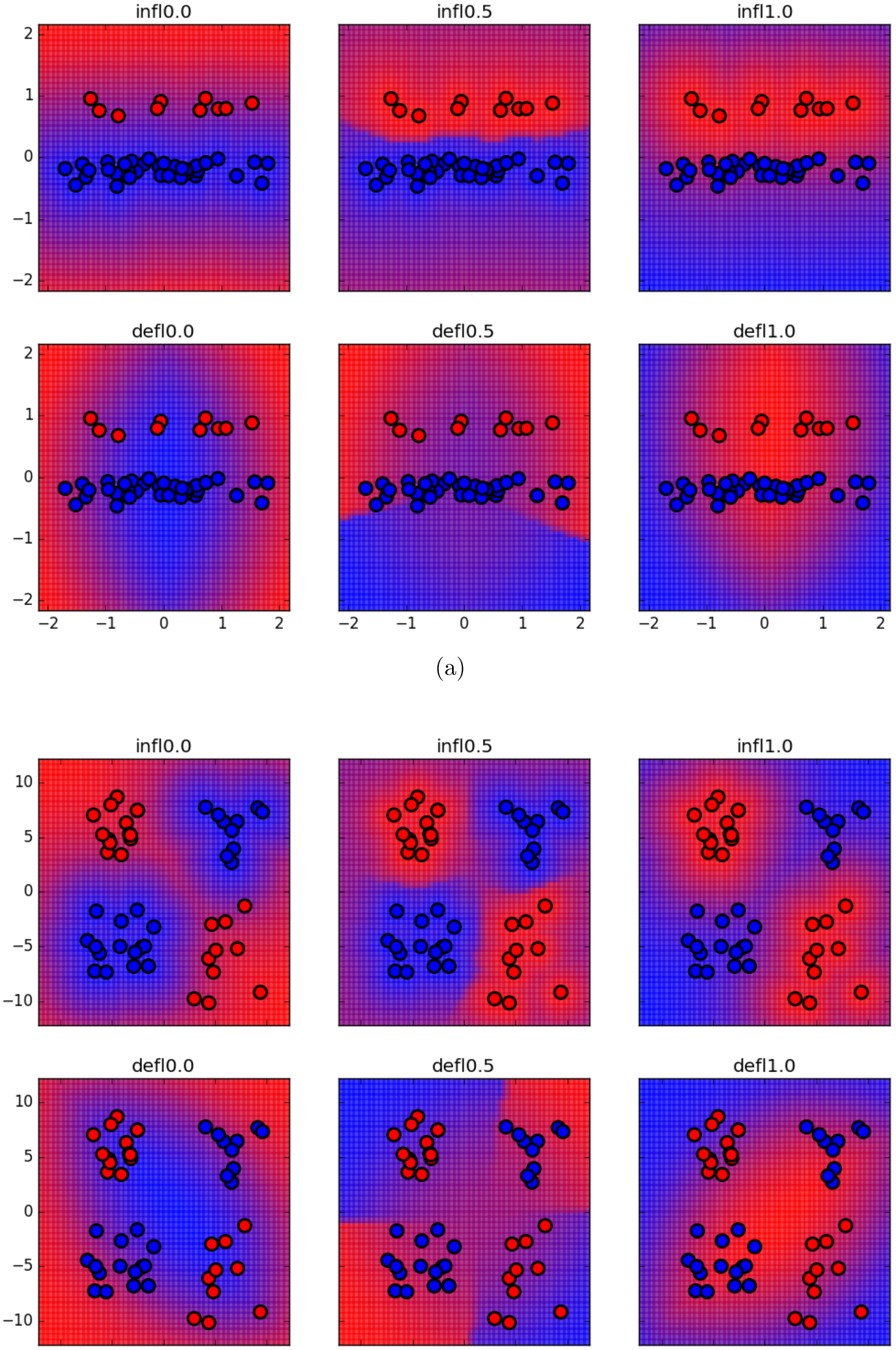
TODO: could generalize this even further to the case where each seed has a different weight

2 Illustrations on 2D Data

In this section we demonstrate the balloon ranking methods on synthetic 2D data[2]. Each set of points contains both positive seeds (red) and negative seeds (blue). The seeds are placed in the center of a grid of linearly discretized points that define the "test" points to be ranked. The inflation and deflation rankings are then computed amongst all of these points using the three settings $\lambda = 0, 0.5, 1$. The grid points are then colored based on their ranking, where red is high confidence positive, blue is high confidence negative, and purple is ambiguous.

These illustrations reveal the topological behavior of the different balloon ranking schemes. In particular, the deflation ranking prioritizes distant points more than nearby points, which can lead to some irregularities. Figure 1b shows how this can prove very bad indeed: in the $\lambda = 0.5$

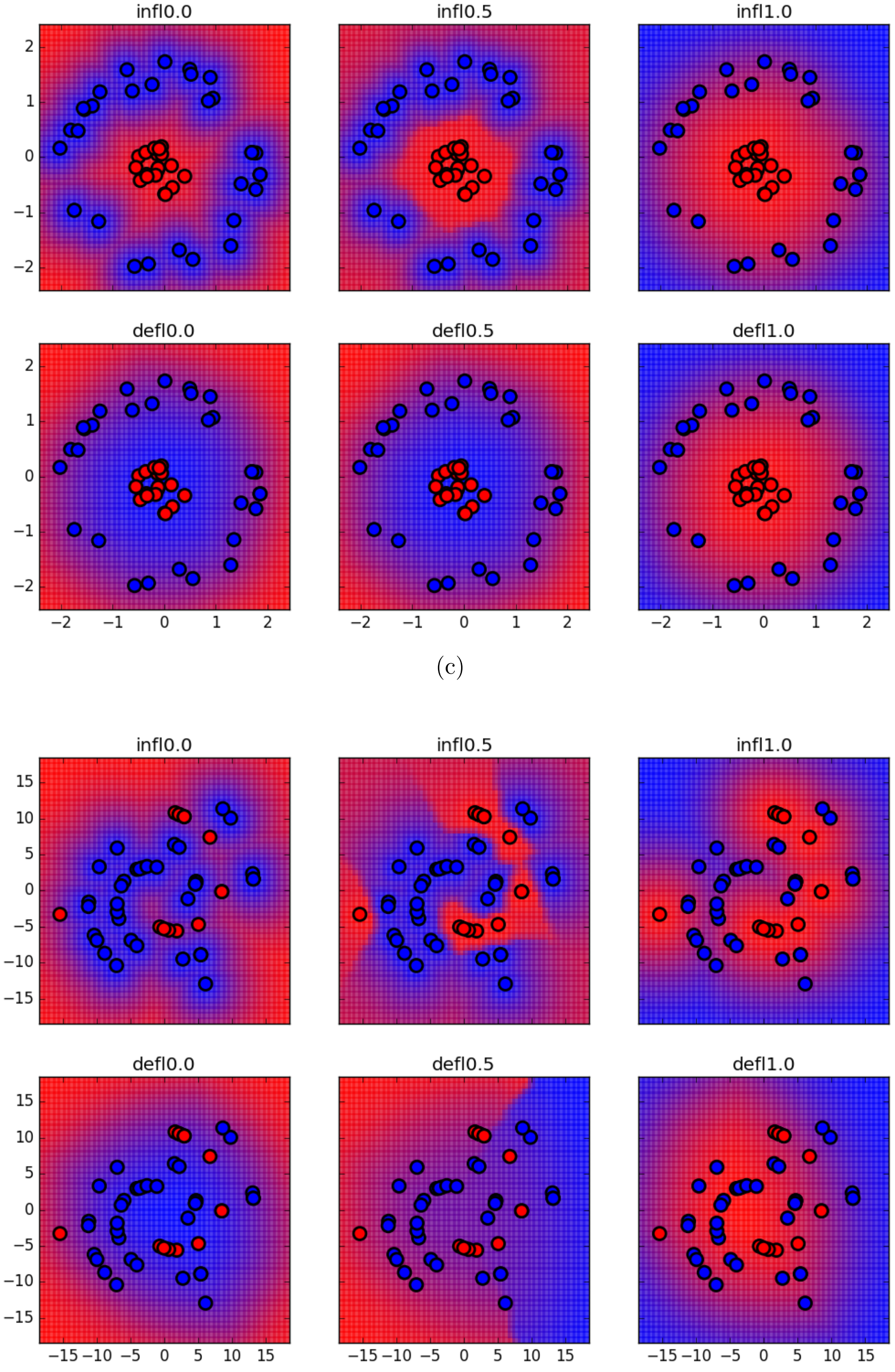
[2] Toy data courtesy https://github.com/deric/clustering-benchmark.



(a)

(b)

Figure 1: Balloon ranking (inflation and deflation) for various settings of λ.



(c)

(d)

Figure 1: Balloon ranking (inflation and deflation) for various settings of λ.

case of deflation, the top-left and bottom-right points will start out with a score of 0, then as deflation occurs, the removal of positive seeds from the opposite corner will cause the ranking to become more negative, even though there is a cluster of positive seeds in the vicinity! A similar phenomenon happens in Figure 1c. The deflation method may therefore be quite problematic when the data is multimodal, or if has regions of one kind of seed enmeshed with or encircling regions of the other kind of seed.

Inflation ranking, on the other hand, prioritizes nearby points in a sensible way so that seeds of one color appear to "radiate" their color as if it were a diffusion process. Depending on λ, seeds of one color or the other will be more responsible for this diffusion. The $\lambda = 0.5$ setting, in particular, produces very "accurate" results on this data set, in that the rankings appear to put high confidence near the respective seed colors, while putting low confidence on points that are roughly equidistant from positive and negative seeds, as well as points that are far from all seeds.