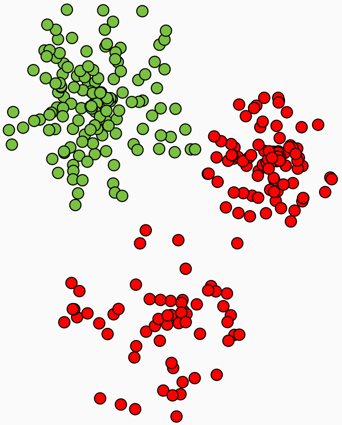
Coresets for Logistic Regression

Chris Schwiegelshohn (joint work with Alexander Munteanu, Christian Sohler, and David Woodruff)

Sapienza University of Rome

Logistic Regression



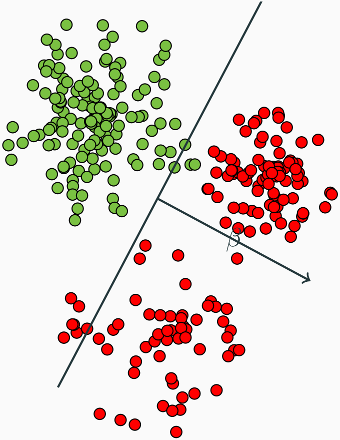
Logistic Regression

Given a point set $X \subset \mathbb{R}^d$, and a labeling function $y : X \to \{-1, 1\}$ find a vector β, such that

$$\sum_{p \in X} \ln(1 + \exp(-y(p) \cdot p^T \beta))$$

is minimized.

Logistic Regression



Logistic Regression

Given a point set $X \subset \mathbb{R}^d$, and a labeling function $y : X \to \{-1, 1\}$ find a vector β, such that

$$\sum_{p \in X} \ln(1 + \exp(-y(p) \cdot p^T \beta))$$

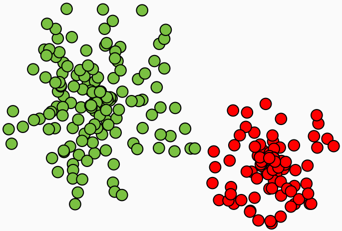
is minimized.

How Can We Summarize This Data Set?

Coreset

Find a set S of points, such that for *any* candidate vector β

$$\text{cost}(X, \beta) \approx \text{cost}(S, \beta).$$

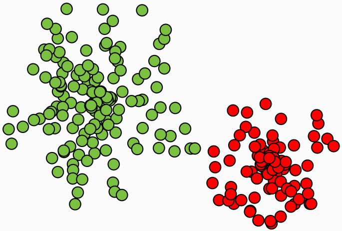


How Can We Summarize This Data Set?

Coreset

Find a set S of points, such that for *any* candidate vector β

$$\text{cost}(X, \beta) \approx \text{cost}(S, \beta).$$

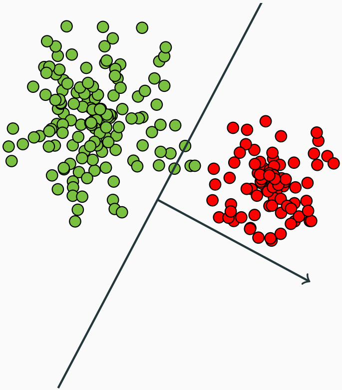


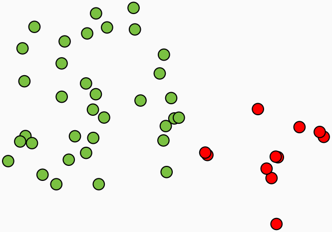
How Can We Summarize This Data Set?

Coreset

Find a set S of points, such that for *any* candidate vector β

$$\text{cost}(X, \beta) \approx \text{cost}(S, \beta).$$



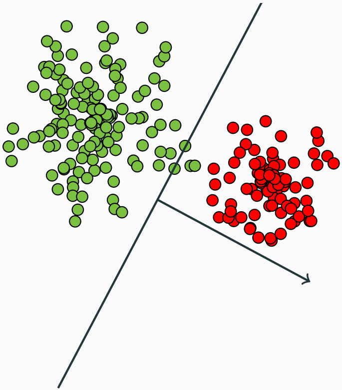


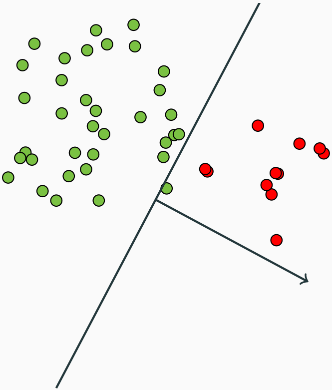
How Can We Summarize This Data Set?

Coreset

Find a set S of points, such that for *any* candidate vector β

$$\text{cost}(X, \beta) \approx \text{cost}(S, \beta).$$



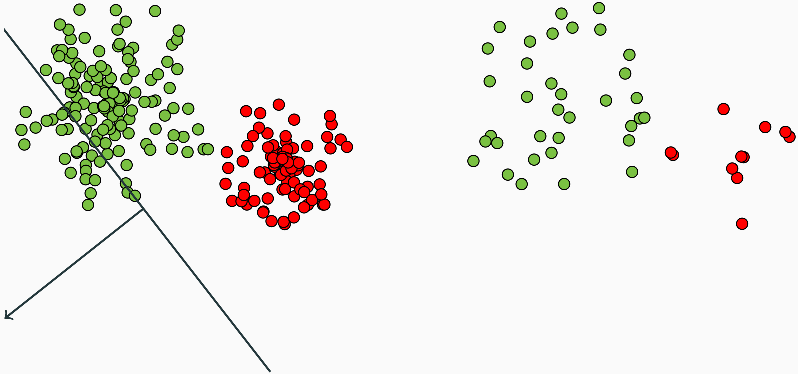


How Can We Summarize This Data Set?

Coreset

Find a set S of points, such that for *any* candidate vector β

$$\mathrm{cost}(X, \beta) \approx \mathrm{cost}(S, \beta).$$

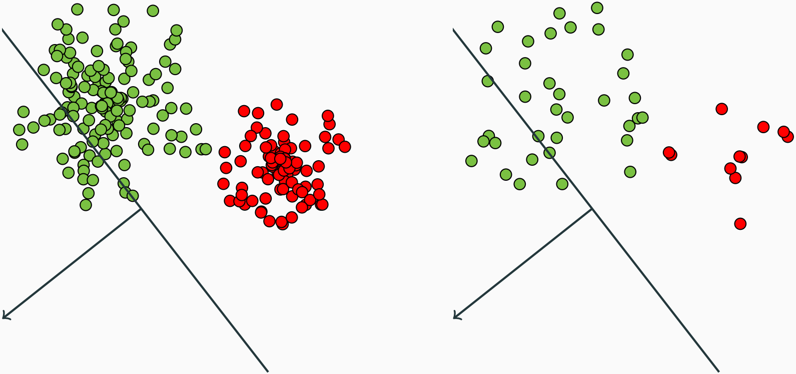


How Can We Summarize This Data Set?

Coreset

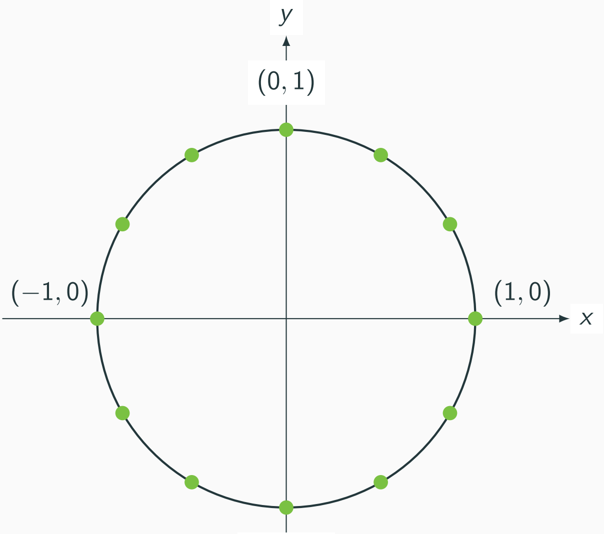
Find a set S of points, such that for *any* candidate vector β

$$\text{cost}(X, \beta) \approx \text{cost}(S, \beta).$$



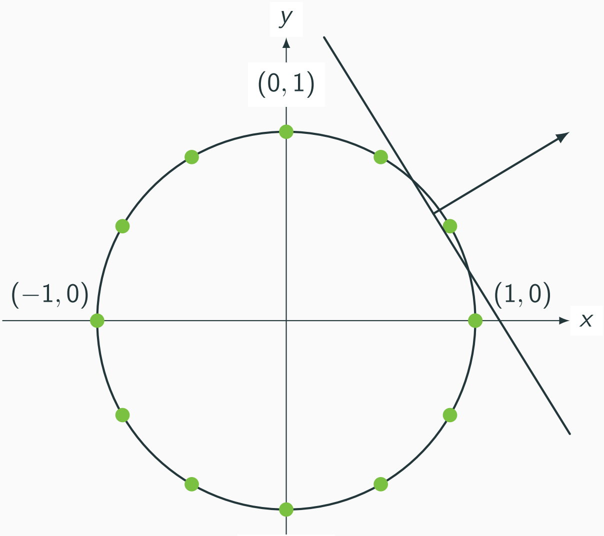
Impossibility Result

$$\text{minimize} \sum_{p \in A} \ln(1 + \exp(-y(p) \cdot p^T \beta))$$



Impossibility Result

$$\text{minimize} \sum_{p \in A} \ln(1 + \exp(-y(p) \cdot p^T \beta))$$



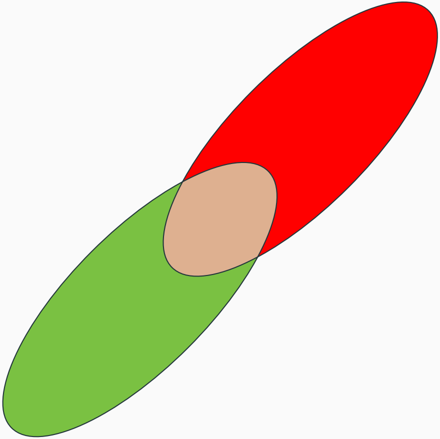
Beyond Worst Case?

Define a notion of overlap μ between the two classes.

Show that the total sensitivity may be bounded in terms of μ.

If μ is large, a suitable sensitivity distribution yields a small coreset.

Works in Streaming, MapReduce, etc.



Coreset Construction via Recursive Sampling

Algorithm

1. Compute $X := U\Sigma V^T$
2. Sample $O(\mu\sqrt{n}\left(\frac{d}{\varepsilon}\right)^2)$ points with replacement with probability proportionate to $\|U_i\|_2$
3. For $i = 1$ to $\log n$
4. Recursively repeat step 2

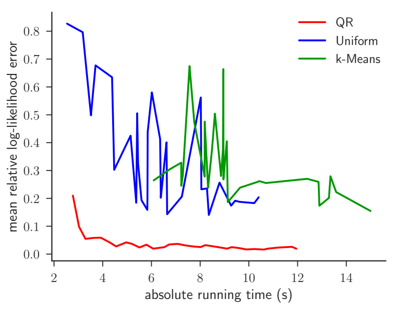
Coreset Construction via Recursive Sampling

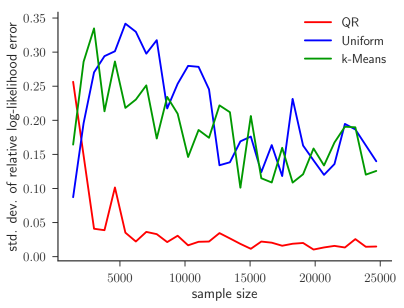
Algorithm

1. Compute $X := U\Sigma V^T$
2. Sample $O(\mu\sqrt{n}\left(\frac{d}{\varepsilon}\right)^2)$ points with replacement with probability proportionate to $\|U_i\|_2$
3. For $i = 1$ to $\log n$
4. Recursively repeat step 2

Algorithm computes a coreset of size $\tilde{O}(\mu^3 d^3 \varepsilon^{-4} \log^4 \mu n d)$.

It Even Works In Practice!





Conclusion and Open Problems

Summary of Results

- Impossibility result for coresets for logistic regression
- Beyond-Worst Case analysis for coreset construction

Open Questions

- Direct sampling scheme that avoids recursion?
- Is μ-complexity the correct measure?
- What other problems admit coresets in "reasonable" cases?