

Contents

1	Introduction and Overview	1
1.1	Organization of the dissertation	6
1.2	Publications relevant to the dissertation	7
1.3	Collaborators	8
2	Practical Coreset Constructions for Machine Learning Problems	9
2.1	Introduction to coresets	10
2.2	Naive approaches to constructing coresets	11
2.3	Importance sampling and sensitivity	12
2.3.1	Uniform guarantee for all queries	15
2.3.2	Main result	16
2.4	Streaming and distributed computation	18
2.5	Coresets for machine learning problems	20
2.5.1	k -means, k -median and Bregman hard clustering	20
2.5.2	Projective clustering, PCA and NNMF	22
2.5.3	Estimation in nonparametric mixture models	23
2.5.4	Classification and regression	24
3	Scalable Training of Mixture Models via Coresets	27
3.1	Motivation and problem statement	28

3.2	Training Gaussian mixture models at scale	33
3.2.1	Strong coresets for λ -semi-spherical Gaussian mixtures	33
3.2.2	Upper bound on the total sensitivity	34
3.2.3	Upper bound on the combinatorial complexity	38
3.2.4	Coreset construction algorithm	40
3.2.5	Sufficient coreset size	42
3.2.6	Fitting Gaussian mixture models	43
3.2.7	Experimental evaluation	45
3.2.8	Other related work	46
3.3	Coresets for soft clustering with regular Bregman divergences	48
3.3.1	Exponential family of distributions	48
3.3.2	Bregman divergences	50
3.3.3	Connection to regular Exponential family mixtures	52
3.3.4	Strong coresets for Bregman soft clustering	55
3.3.5	Upper bound on the total sensitivity	56
3.3.6	Upper bound on the combinatorial complexity	59
3.3.7	Coreset construction algorithm	62
3.3.8	Sufficient coreset size	64
3.3.9	Soft Bregman clustering	65
3.3.10	Gaussian mixture models and Bregman divergences	65
3.3.11	Experimental evaluation	66
3.3.12	Other related work	68
4	Linear-time Outlier Detection via Sensitivity	69
4.1	Motivation and problem statement	70
4.2	Distance-based outlier detection	71
4.3	Sensitivity as a measure of outlierness	73
4.3.1	Sensitivity and clustering quality	73

4.3.2	Lower-bound on the sensitivity	74
4.3.3	A natural interpretation of influence	75
4.3.4	Generality and robustness	76
4.4	Efficient computation of influence	76
4.5	Experimental evaluation	78
5	Tradeoffs for Space, Time, Data and Risk in Unsupervised Learning	83
5.1	Motivation and problem statement	84
5.2	The statistical k -means problem	86
5.3	Data summarization	87
5.4	Space-time-data-risk tradeoff	88
5.5	Existence of tradeoffs	91
5.5.1	Sufficient conditions	92
5.5.2	Risk bounds	95
5.5.3	Running time bounds	97
5.5.4	Verification of tradeoffs	97
5.5.5	Illustration of the main theorem and resulting tradeoffs	98
5.6	Data-driven tradeoff navigation	100
5.6.1	Theoretical setting	100
5.6.2	A tradeoff navigation algorithm (TRAM)	101
5.7	Experimental evaluation	107
6	Open Problems and Future Work	113