Lecture I – Big Data in Machine Learning

Marina Meilă

Department of Statistics University of Washington

STAT 548/CSE 547 Spring, 2025

イロン イ団 とく ヨン イヨン

STAT 548/CSE 547Spring, 2025

2

1/

Marina Meilă (Statistics)

Big data and Machine Learning I

Big Data has implications for ML at many levels

- Storage
 - may not fit in local memory
 - expensive/slow to move around
 - I/O expensive/slow
- Access
 - serial/by block, not random
- Indexing
 - · Preprocessing steps that allow faster access during
- Computing
 - Parallelization when possible
 - Automation of resource management (Hadoop, Spark)
- Algorithms
 - predominantly sub-quadratic, i.e. $\mathcal{O}(n), \tilde{\mathcal{O}}(n)$
 - sub-linear, i.e. o(n) when possible sampling, Stochastic Gradient Descent (SGD)

イロト 不得 トイヨト イヨト

STAT 548/CSE 547Spring, 2025

3

2/

Big data and Machine Learning II

- Tasks
 - streaming
 - bandits
 - on-line learning
 - approximate rather than exact solutions (e.g. nearest-neighbors)
- Statistical
 - new problems (streaming, bandits)
 - what is i.i.d. sampling anyways? (on-line learning)
 - approximation and sampling (e.g. how to sample from a data stream)
 - can ask more detailed questions non-parametric statistics
 - more spurious patterns to find validation without human intervention

イロン イ団 とく ヨン イヨン

STAT 548/CSE 547Spring, 2025

э

3/

- often high dimension D
- Solves curse of dimensionality? No.

Parametric vs. non-parametric

A mathematical definition

• A model class \mathcal{F} is parametric if it is finite-dimensional, otherwise it is non-parametric

In other words

- When we estimate a parametric model from data, there is a fixed number of parameters, (you can think of them as one for each dimension, although this is not always true), that we need to estimate to obtain an estimate $\hat{f} \in \mathcal{F}$.
- The parameters are meaningful.
 E.g. the β_j in logistic regression has a precise meaning: the component of the normal to the decision boundary along coordinate j.

イロト 不得 トイヨト イヨト

STAT 548/CSE 547Spring, 2025

3

4 /

• The dimension of β does not change if the sample size *n* increases.

Non-parametric models - Some intuition

- \bullet When the model is non-parametric, the model class ${\cal F}$ is a function space .
- The \hat{f} that we estimate will depend on some numerical values (and we could call them parameters), but these values have little meaning taken individually .
- The number of values needed to describe \hat{f} generally grows with *n*. Examples In the Nearest neighbor and kernel predictors, we have to store a II the data points, thus the number of values describing the predictor *f* grows (linearly) with the sample size. Exercise Does the number of values describing *f* always grow linearly with the sample size? Does it have to always grow to infinity? Doe s it have to always grow in the same way for a given \mathcal{F} ?
- Non-parametric models often have a smoothness parameter. Examples of smoothness parameters K in K-nearest neighbor, h the kernel bandwidth in kernel regression.

To make matters worse, a smoothness parameter is not a parameter! More p recisely it is not a parameter of an $f \in \mathcal{F}$, because it is not estimated from the data, but a descriptor of the model class \mathcal{F} .

・ロト ・ 日 ト ・ 日 ト ・ 日 ト ・

STAT 548/CSE 547Spring, 2025

5/

• We will return to smoothness parameters later in this lecture.

Parametric vs. non-parametric models

Parametric

- Linear, logistic regression
- Linear Discriminant Analysis (LDA)
- Neural networks (if not very large)
- Naive Bayes
- CART with L levels
- Clustering by k-means, finite mixture models
- Spectral clustering of graphs

Non-parametric

- Nearest-neighbor classifiers and regressors
- Kernel (e.g. Nataraya-Watson) regression
- Monotonic regression
- Support Vector Machines
- Large, overparametrized neural networks
- Dirichlet Process Mixture Models (DPMM)
- Clustering by level sets or mode-finding

イロト 不得 トイヨト イヨト

STAT 548/CSE 547Spring, 2025

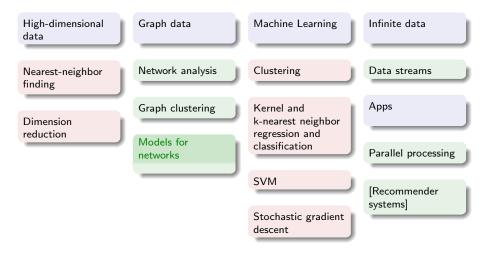
э

6/

- Affinity based clustering
- Manifold learning

This course

This course



STAT 548/CSE 547Spring, 2025 7/7

э

・ロト ・四ト ・ヨト ・ヨト