Advanced Machine Learning, 13 Aug 2019

Sourish Das, Madhavan Mukund

Deep Learning

- Acyclic NN (feed forward)
  - Loss functions, Optimization Strategies, Regularization
- Convolutional NN
- Recurrent NN, LSTM

Reinforcement Learning

└ AlphaGo

Probabilistic Graphical Models

Assignments ~35%

Mid Sem Exam ~25%

Final Exam ~40%

..

Motivation

ML – Builds a model of a "concept" from training examples

Concept?

Data Space $\quad D \quad \langle x_1, x_2, \ldots, x_n \rangle$

Concept $\quad C \subseteq D$

Build a model – construct a description of C from training data

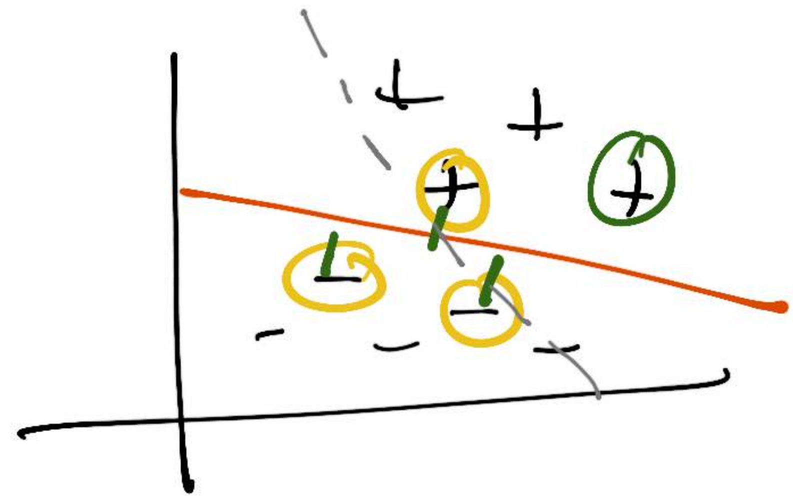
Regression

$$y = f(x_1, \ldots, x_n)$$
$$= \sum w_i x_i + b$$

Optimize w_i's

Cost/Loss associated with w_i's

SVM



Optimize min distance to bounding point

What distinguishes ML from optimization?

Challenge is <u>generalization</u>

How well the model performs on <u>unseen</u> data

Prima facie - impossible

Unseen data is "similar" to training data

Probabilistic setting

- Distribution over D - p
- Training & Unknown samples are i.i.d wrt p

D - all data points

$C \subseteq D$ concept

P - distribution over D

Space of possible models

$\mathcal{H}$ - hypothesis space

e.g regression or SVM

- all possible hyperplanes

PAC Learning (Leslie Valiant)

Probabilistically Approximately Correct

Fix p, D, c, $\mathcal{H}$

Connect error on training data

(optimization while constructing $h \in \mathcal{H}$)

to potential error on unknown (general) data

Larger training data $\Rightarrow$ Better h

Given $p, D, C, \mathcal{H}$

fix parameters δ, ε

Can compute an n (as a function of δ, ε)

s.t. that any h built from a

training sample of size n

With probability $1-\delta$

if h is ε-away from training data,

h is also ε-away from unseen data

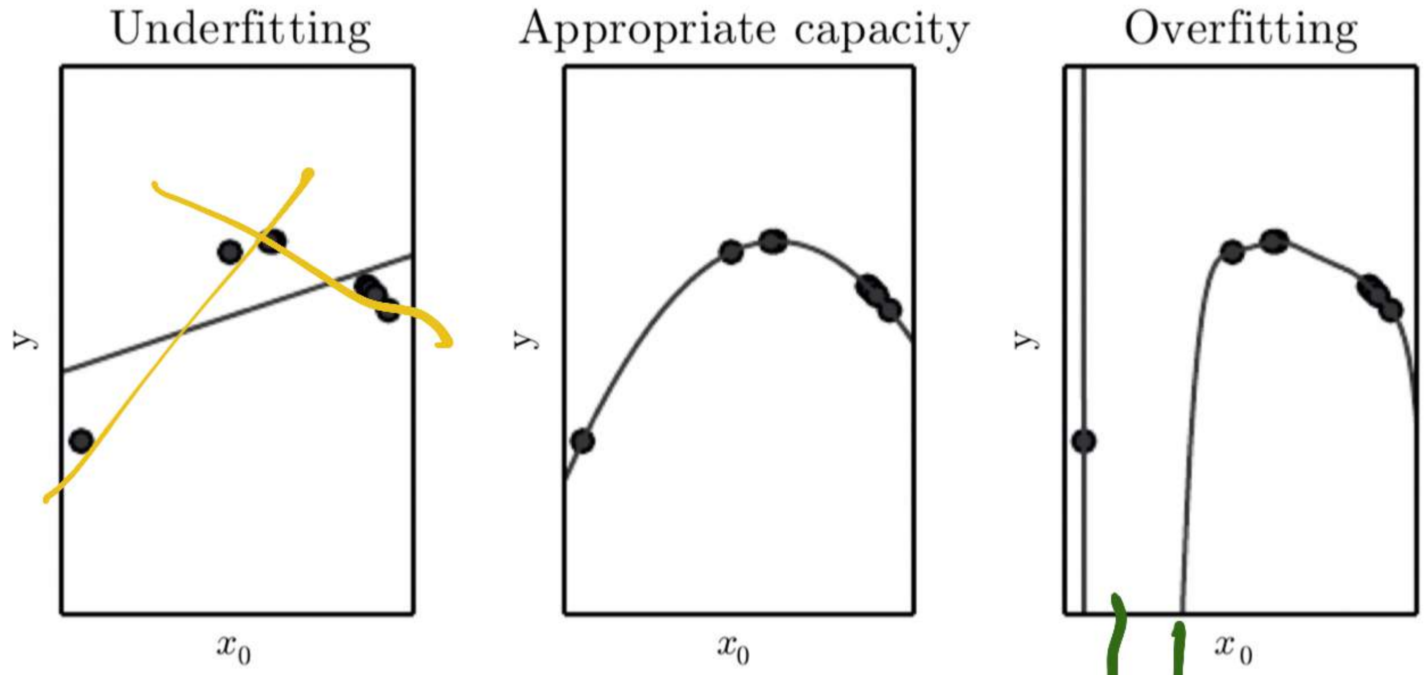
Expressiveness of $\mathcal{H}$ — Capacity of model

Statistical ML → Vapnik-Chervonenkis (VC) Dimension

Regression

$$w_1 x + b \rightarrow w_1 x + w_2 x^2 + b$$

$$\ldots \quad w_1 x \ldots + w_k x^k + b$$



Low
Capacity

Match expressiveness of $\mathcal{H}$ to

complexity of C

Occam's Razor - Simplest model that fits is the best

Penalize complex models in $\mathcal{H}$

Polynomial: $\sum_{i=0}^{k} w_i x^i$ $\quad W = \langle w_0, \ldots, w_k \rangle$

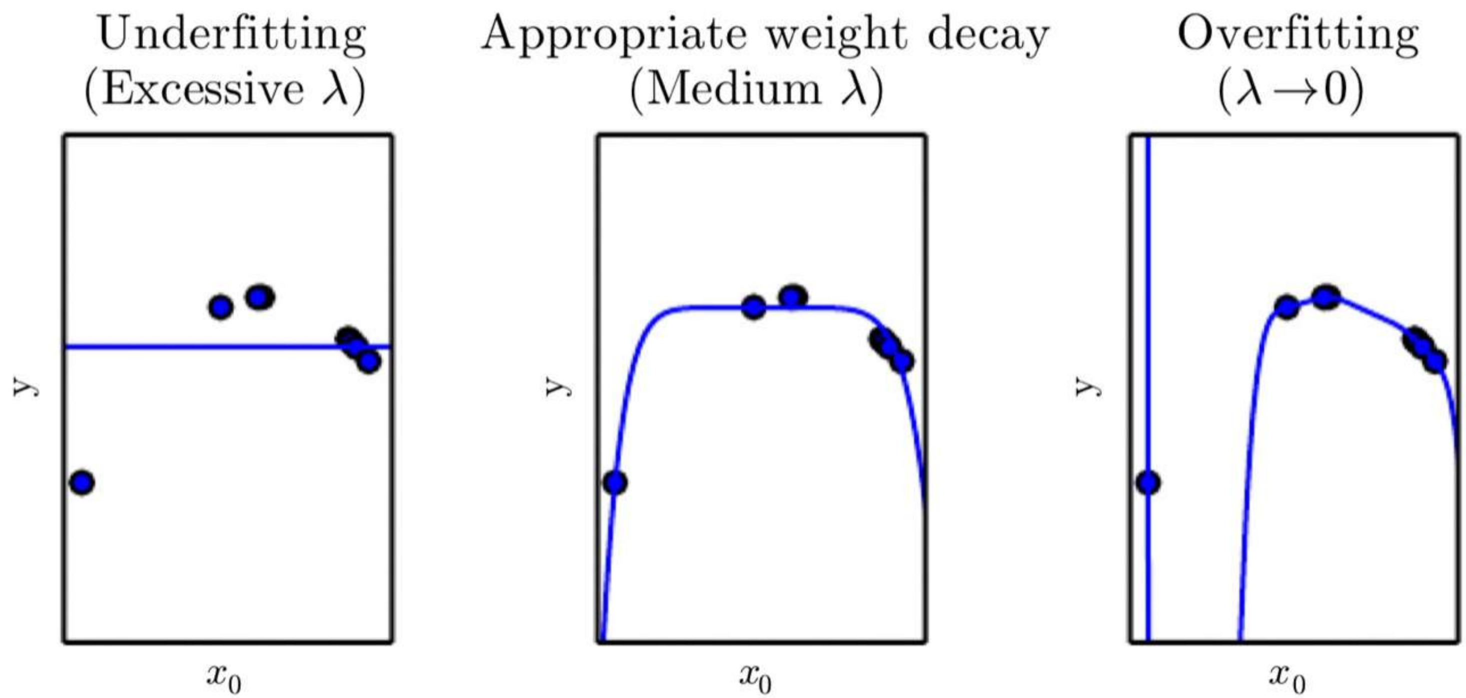
Add a cost proportional to w_i's

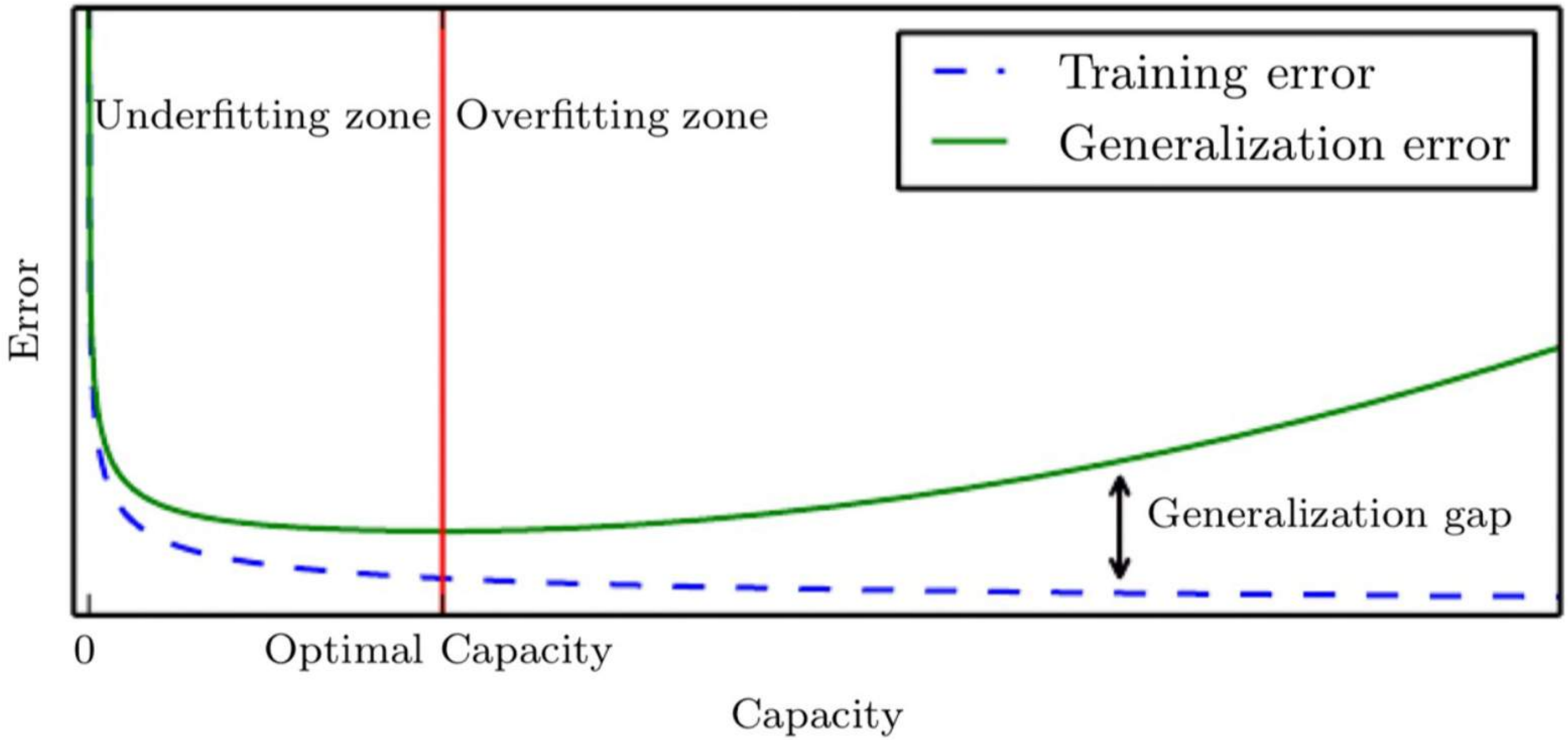
Typical $W^T W$

Optimize: Original loss + $\lambda \cdot$ Complexity

Regularization

Regularization is any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error.





Regularization tries to push the hypothesis towards the red line from the right

Hyperparameters

ML estimates parameters of a given h

Degree of polynomial ?

Don't want ML to "discover" this

Freeze a prior - Hyperparameter

Another Example - regularization coefficient

λ

Bias & Variance

Bias – Expressiveness / Ability to describe C

Variance – Sensitivity to training data

Statistically

Across all training data : $E(C)$

Real C

"Difference" $E(C) - C$ = bias

Variance

Statistical variance between $E(C)$ and a given model h

Neural Networks

Networks of linear separators

SVM – non linear data via geometric transforms – "Kernel trick"

Network of perceptrons has (arbitrarily) high capacity

Applying regularization &

optimization (variations on

gradient descent)