Machine Learning: Basic Principles

Model Validation and Selection

Salo, September 2018

Guiding Motto

never fall in love with your favourite model!

in this lecture "model" $=$ hypothesis space $\mathcal{H}$ (which is a subset of all mappings $h(\cdot) : \mathcal{X} \rightarrow \mathcal{Y}$) ;-)

Background

this lecture is inspired by

- lecture notes
  http://cs229.stanford.edu/notes/cs229-notes5.pdf
  of Prof. Ng (Stanford)
- video of Prof. Ng
  https://www.youtube.com/watch?v=MyBSkmUeIEs
- Chapter 5.3 of the "deep learning book"
  http://www.deeplearningbook.org

Outline

1. Intro
2. A Simple Model Selection Method
3. Wrap Up

Ski Resort Marketing

- you still did not find another job
- thus, you still work as marketing of a ski resort
- hard disk full of webcam snapshots (gigabytes of data)
- you want order them according to daytime of snapshots
- you have only a few hours for this task ...

A Webcam Snapshot

a data point = a single webcam snapshot

feature vector given by
green intensity for EACH pixel

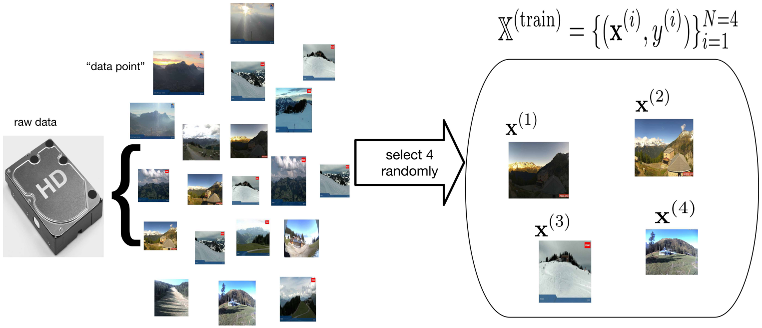
label/target/output y

ML workflow so far...

- create dataset $\mathbb{X}^{(\mathrm{train})} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$ by manual labeling
- features $\mathbf{x}^{(i)} \in \mathcal{X}$ and label $y^{(i)} \in \mathcal{Y}$ of ith data point
- define loss $L((\mathbf{x}, y), h(\cdot))$ (e.g., $L((\mathbf{x}, y), h(\cdot)) = (y - h(\mathbf{x}))^2$)
- define hypothesis space $\mathcal{H}$ (e.g., linear maps $h(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$)
- learn predictor $h(\cdot) : \mathcal{X} \rightarrow \mathcal{Y}$ by empirical risk minimization

$$\min_{h(\cdot) \in \mathcal{H}} \mathcal{E}\{h(\cdot) | \mathbb{X}^{(\mathrm{train})}\} = (1/N) \sum_{i=1}^{N} L((\mathbf{x}^{(i)}, y^{(i)}), h(\cdot))$$

The Dataset

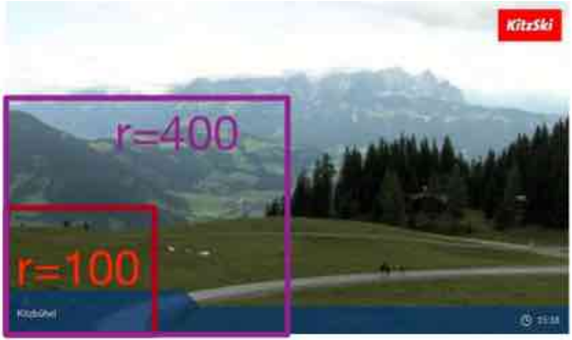

"data point"
raw data
HD
select 4
randomly
$\mathbb{X}^{(\text{train})} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N=4}$
$\mathbf{x}^{(1)}$
$\mathbf{x}^{(2)}$
$\mathbf{x}^{(3)}$
$\mathbf{x}^{(4)}$

The Features

- we assume that all images consist of d pixels
- represent a snapshot by vector $\mathbf{x} \in \mathbb{R}^d$
- individual feature x_i represents green level of pixel i
- lets collect all pixels i in the lower left square of size r into

$$\mathcal{R}_r = \{ \text{ pixels in the lower left square of size } r \text{ pixels} \}$$

Lower Left Squares



use bottom left square with r pixels

The Hypothesis Space

- we predict daytime y using a linear map $h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$
- weight vector $\mathbf{w} \in \mathbb{R}^d$ long for typical image sizes
- consider subset of mappings (hypothesis space)

$$\mathcal{H}^{(r)} = \{h(\mathbf{x}) = \mathbf{w}^T\mathbf{x} : w_i = 0 \text{ for } i \notin \mathcal{R}_r\}$$

- $\mathcal{H}^{(r)}$ contains linear maps from pixels $\mathbf{x} \in \mathbb{R}^d$ to predicted label $\hat{y} = h(\mathbf{x})$ which take only pixels in $\mathcal{R}_r$ into account

The Empirical Risk Minimization

- consider a predictor $h^{(\mathbf{w})} \in \mathcal{H}^{(r)}$
- prediction incurs loss (error) $L((\mathbf{x}, y), h(\cdot)) = (y - h(\mathbf{x}))^2$
- empirical risk $\mathcal{E}\{h^{(\mathbf{w})} | \mathbb{X}^{(\mathrm{train})}\}$ =average loss on $\mathbb{X}^{(\mathrm{train})}$
- for a particular model $\mathcal{H}^{(r)}$, choose optimal $\mathbf{w}_r$ via ERM

$$\mathbf{w}_r = \underset{\mathbf{w}: h^{(\mathbf{w})} \in \mathcal{H}^{(r)}}{\operatorname{argmin}} (1/N) \sum_{i=1}^{N} (y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2$$

$$= \underset{\mathbf{w}: w_i = 0 \forall i \notin \mathcal{R}_r}{\operatorname{argmin}} (1/N) \sum_{i=1}^{N} (y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2$$

The Million Dollar Question

- which hypothesis space (model) $\mathcal{H}^{(r)}$ should we use ?
- what is the best choice for the model parameter r ?
- r is the number of pixels used for predicting daytime

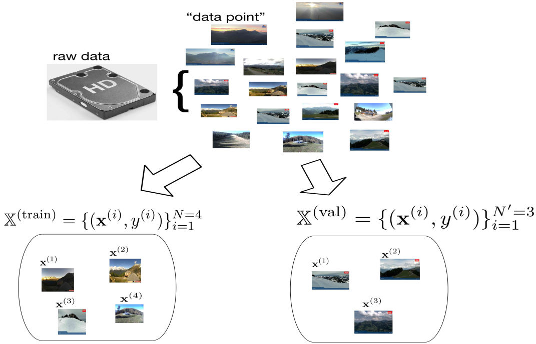
Outline

A First Shot...

- lets try out ERM with $\mathcal{H}^{(r)}$ for different choices of r
- for each value r, get optimal predictor $h^{(\mathbf{w}_r)}(\mathbf{x}) = \mathbf{w}_r^T \mathbf{x}$
- choose r yielding smallest training error $\mathcal{E}\{h^{(\mathbf{w}_r)}|\mathbb{X}^{(\mathrm{train})}\}$
- THIS WILL NOT WORK!

The Training Error vs. Model Size

Use Different Data for Training and Validation



1. ERM on dataset $\mathbb{X}^{(\mathrm{train})}$ to find optimal predictor $h^{(\mathbf{w}_r)}(\cdot)$
2. apply $h^{(\mathbf{w}_r)}(\cdot)$ to another dataset $\mathbb{X}^{(\mathrm{val})}$ to get average loss

$$(1/N') \sum_{(\mathbf{x},y)\in\mathbb{X}^{(\mathrm{val})}} L((\mathbf{x}, y), h_{\mathrm{opt}}(\cdot))$$

Training and Validation Error vs. Model Size r

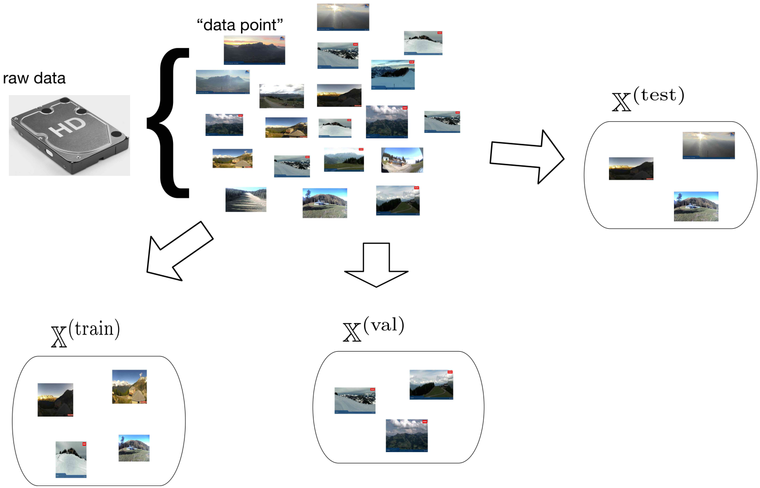
A Simple Model Selection Method

- lets try out ERM with $\mathcal{H}^{(r)}$ for different choices of r
- for each value r, get optimal predictor $h^{(\mathbf{w}_r)}(\mathbf{x}) = \mathbf{w}_r^T \mathbf{x}$
- choose $r = r'$ yielding smallest validation error $\mathcal{E}\{h^{(\mathbf{w}_r)}|\mathbb{X}^{(\mathrm{val})}\}$
- THIS WILL WORK!

Validating the Final Model

- how to validate he finally selected predictor $h^{(\mathbf{w}_{r'})}(\mathbf{x})$?
- can we use validation error $\mathcal{E}\{h^{(\mathbf{w}_{r'})}|\mathbb{X}^{(\mathrm{val})}\}$?
- we have used $\mathbb{X}^{(\mathrm{val})}$ to learn (choose) the optimal r !
- thus we need one further dataset, the test set $\mathbb{X}^{(\mathrm{test})}$

The Dataset


raw data
HD
“data point”
$\mathbb{X}^{(\text{test})}$
$\mathbb{X}^{(\text{train})}$
$\mathbb{X}^{(\text{val})}$

A Simple Model Selection Method

- generate different sets of labeled data $\mathbb{X}^{(\mathrm{train})}, \mathbb{X}^{(\mathrm{val})}, \mathbb{X}^{(\mathrm{test})}$
- find optimal predictor (via ERM on $\mathbb{X}^{(\mathrm{train})}$) for $\mathcal{H}^{(r)}$ using different choices of r
- for each value r, another optimal predictor $h^{(\mathbf{w}_r)}(\mathbf{x}) = \mathbf{w}_r^T \mathbf{x}$
- choose $r = r'$ yielding smallest validation error $\mathcal{E}\{h^{(\mathbf{w}_r)} | \mathbb{X}^{(\mathrm{val})}\}$
- evaluate final predictor using error on test set $\mathcal{E}\{h^{(\mathbf{w}_{r'})} | \mathbb{X}^{(\mathrm{test})}\}$

Outline

1. Intro
2. A Simple Model Selection Method
3. Wrap Up

A Golden Rule of ML Practice

- for given model (hypothesis space) use ERM on training set
- compute validation error of optimal predictor on validation set
- choose best model according to validation error
- evaluate optimal predictor within best model using test set