Lecture 1 - Introduction

Advanced course in Statistical Machine Learning: Theory and Applications

Stéphane Canu

scanu@axiom.anu.edu.au.

National ICT of Australia
and
Australian National University.

Thanks to Alex Smola and S.V.N. "Vishy" Vishwanathan for intial version of slides.

Statistical machine learning?

Definition : the ability of a machine to (automatically) improve its performance based on previous results

A new way of programming:

- tell the machine what to do
- show the machine behaviors (watch what I do)

Henry Lieberman - `http://web.media.mit.edu/~lieber/PBE/`

Statistics with:

- large data sets — computational issues
- no specific model — model selection issues

Programming through examples: what's new?

Machine learning problems

– Learning for what?

- moving, planning
- speech
- writting
- language, translation
- vision,

– Why is it difficult?

1. size effect
   - sample size - number of examples
   - dimensionality - size of each example
2. unknown model (non linear)
3. computing complexity

Machine learning promises

- Short term: help us with some tedious task — show patterns
- Mid term: interact with human (HCI) — show behaviors
- Long term: understand the nature of information — ???

How to deal with these questions

4 Tipical problems

Tasks

1. Optical character recognition (OCR)
2. query Google
3. Face recognition
4. DNA language

What for:

- Help human	Engineering
- H. C. I.	Computer Science
- Knowledge - information	Maths

On line learning vs. batch learning

Learning: that is the question

Notations: input $x \in \mathcal{X}$ (the domaine), output $y \in \mathcal{Y}$ (the codomaine)

Task	*Input*	*Output*	*Cost*
OCR	16 $\times$ 16	$\{0, 1, \cdots, 9\}$	$0/1$ or $
Google	query	web pages	are you happy!
Face recognition	image	mummy!	$0/1$
DNA language	microarray	genes	probability

Prediction problem I (very vague version)

$$\boxed{\text{find } f : \mathcal{X} \longrightarrow \mathcal{Y} \text{ such that } f(x) \text{ looks like } y}$$

Probabilistic setup

- X is a random variable in $\mathcal{X} \subset \mathbb{R}^d$ (d is refered as the dimensionality)
- Y is a random variable in $\mathcal{Y}$,
- (X, Y) follows $\mathbb{P}(x, y) \in \mathcal{P} \cdots$ $\boxed{\text{unknown!}}$

we are learning in a probabilistic framework

Learning: what is given?

Given information

- Data: a sample $S_n = (X_i, Y_i)_{i=1,n}$, drawn according to $\mathbb{P}(x, y)$ still unknown,
- Cost function (loss) $C \; : \; (\mathcal{X}, \mathcal{Y}, \mathcal{Y}) \longrightarrow \mathbb{R}$,
- Prior information about the nature of the solution.

Prediction problem II (more precise but unfeasible)

$$\boxed{\text{find } f^* : \mathcal{X} \longrightarrow \mathcal{Y} \text{ minimizing } J(f) \stackrel{\Delta}{=} \mathbb{E}(C(X, Y, f(X)))}$$

Use data!

Prediction problem III (feasible but to be made precise)

$$\boxed{\text{estimate } f^*}$$

what can we do with the data?

Two ways to deal with the data

the empirical cost

$$
\begin{aligned}
J(f) &= \mathbb{E}(C(X,Y,f(X)) \\
&= \int \dots \int \; C(\mathbf{x},y,f(\mathbf{x}))\; \mathbb{P}(\mathbf{x},y)\; d\mathbf{x}dy
\end{aligned}
$$

$$
\begin{aligned}
J_{\text{emp}}(f) &= \text{Mean}(C(X,Y,f(X)) \\
&= \int \dots \int \; C(\mathbf{x},y,f(\mathbf{x}))\; \mathbb{P}_{\text{emp}}(\mathbf{x},y)\; d\mathbf{x}dy \\
&= \frac{1}{n}\sum_{i=1}^{n} C(\mathbf{x}_i,y_i,f(\mathbf{x}_i))
\end{aligned}
$$

the likelihood for some parameters $\theta : \mathbb{P}(x,y) = g(x,y,\theta)$

$$
\begin{aligned}
\mathcal{L}(X_i,Y_i,\theta) &= -\log \mathbb{P}(x_1,y_2,\cdots,x_i,y_i,\cdots,x_n,y_n|\theta) \\
&= \sum_{i=1}^{n} -\log \mathbb{P}(x_i,y_i|\theta)
\end{aligned}
$$

Minimize the cost or maximize the likelihood

An interesting particular case

the least square $C(f) = \|f(\mathbf{x}) - y\|^2$

$$J_{\text{emp}}(f) = \frac{1}{n}\sum_{i=1}^{n} \|f(\mathbf{x}_i) - y_i\|^2$$

the gausian case: $\mathbb{P}(x, y, \theta) = \frac{1}{Z(\sigma)} \exp^{-\frac{1}{2\sigma}\|f(\mathbf{x}_i) - y_i\|^2}$, $\theta = (f, \sigma)$

$$\begin{aligned} \mathcal{L}(X_i, Y_i, \theta) &= \sum_{i=1}^{n} -\log \mathbb{P}(x_i, y_i|\theta) \\ &= \frac{1}{2\sigma}\sum_{i=1}^{n} \|f(\mathbf{x}_i) - y_i\|^2 - n \log Z(\sigma) \end{aligned}$$

the target function (for a given $\mathbf{x}$) $\min_p \mathbb{E}(p - Y)^2$

$$\begin{aligned} \mathbb{E}(p - Y)^2 &= \int (p - y)^2 \mathbb{P}(y|\mathbf{x}) dy \\ &= p^2 - 2p \underbrace{\int y \mathbb{P}(y|\mathbf{x}) dy}_{f^*(x) = \mathbb{E}(Y|\mathbf{x})} + \int y^2 \mathbb{P}(y|\mathbf{x}) dy \end{aligned}$$

different approaches can lead to *analogous* algorithms

Exercice

What is the target function if the cost is (for a given $\mathbf{x}$)

$$C(\mathbf{x}, y, p) = y \log p + (1 - y) \log(1 - p)$$

$$C(\mathbf{x}, y, p) = \left(\frac{p - y}{y}\right)^2$$

$$C(\mathbf{x}, y, p) = |p - y|$$

Two learning strategies

Strategy I: Structural risk minimization

- choose a structure for f (hypothesis class)
- prove $J < J_{\text{emp}} + \text{bound}$
- minimize J_{emp} and the bound

Strategy II: Model and conquer

- model $\mathbb{P}(x, y)$ — exponential family
- deduce a feasible criterion — some specific work
- minimize it — optimization issues

Both strategies can lead to the same algorithms

what is the best?

- no method is “silver bullet”
- no method is universally better

  no free lunch
- each method has its own fitted data

Data Set Selection - `http://www.jmlg.org/papers.htm`

- how to choose?
  - use prior knowledge
  - make assumptions
  - choose a stable or robust one (if they are wrong)

Jerome H. Friedman - `http://www.stanford.edu/class/stats315b/`

If possible, try several - estimate best or use committee

Course overview

Advanced course in Statistical Machine Learning: Theory and Applications

1. Mathematical background
2. Density estimation and Exponential family
3. Kernels
4. Supevised learning I
5. Supevised learning II
6. Graphical models I
7. Graphical models II
8. Conditional random fields
9. Boosting

Focus on exponential family & algorithm

References

- Statistical learning theory
  - Kernel Machines, 2002
  - Vladimir Vapnik. The Nature of Statistical Learning Theory. Springer, 1995.
  - Nello Cristianini and John Shawe-Taylor. An Introduction to Support Vector Machines. Cambridge University Press, Cambridge, UK, 2000
- Statistical patern recognition
  - R. O. Duda, P. E. Hart and D. G. Stork, Pattern Classification (2nd ed.), John Wiley and Sons, 2001.
  - T. Hastie, R. Tibshirani, and J. Fridman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, 2001
  - D. Hand, H. Mannila, P. Smyth, Principles of Data Mining, MIT Press, 2001
- ...On the web
  - http://kernel-machines.org
  - http://www.ph.tn.tudelft.nl/PRInfo/
  - http://citeseer.nj.nec.com/
  - http://asi.insa-rouen.fr/~scanu

Questions?