Introduction to Statistical Learning #6

Nicolas Vayatis

CMLA - ENS Cachan

MVA 2016-2017

Course outline

- Chapter 1 - Modeling aspects:
  - Data, decision, risk, optimality
- Chapter 2 - Tools:
  - Concentration inequalities and complexity measures
- Chapter 3 - Theory:
  - Consistency and error bounds of mainstream learning algorithms
- Chapter 4 - Advanced topics:
  - Multiclass classification, structured prediction, sequential learning, ...

End of the course

- Today (Dec. 6) $\rightarrow$ last theoretical class
- Grades on partial exam available by Dec. 20
- On Dec. 13 $\rightarrow$ last exercise session
- Final exam on Friday Jan. 6 from 2pm to 4pm (with documents)

Chapter 4 - Advanced topics

Overview of Chapter 4 - Advanced topics

1. Multiclass classification
2. Structured prediction $\Rightarrow$ J.P. Vert-J. Mairal's course
3. Sequential learning $\Rightarrow$ V. Perchet's course
4. Further topics: Ranking, Metric learning, Unsupervised/Semisupervised learning, Transfer learning, Multiview learning, Link prediction, ...

Chapter 4 - Advanced topics

Multiclass classification

Multiclass classification

- Focus on monolabel case: $Y \in \{1, \ldots, K\}$ (see also multilabel $\{-1, +1\}^K$)
- Risk criterion (generic) of a multiclass classifier g:

$$L(g) = \mathbb{P}(Y \neq g(X)) \ , \qquad \text{(Hamming distance)}$$

- Scoring approach to multiclass classification: real-valued $h(x, y)$
- Margin concept:

$$\rho_h(x, y) = h(x, y) - \max_{y' \neq y} h(x, y')$$

Rademacher average

- Assume we have classes of decision functions $\mathcal{F}_1, \ldots, \mathcal{F}_M$
- Consider the class

$$\mathcal{G} = \{\max(h_1, \ldots, h_M) \ : \ h_m \in \mathcal{F}_m\}$$

- Then:

$$\widehat{R}_n(\mathcal{G}) \leq \sum_{m=1}^{M} \widehat{R}_n(\mathcal{F}_m)$$

Margin bound for multiclass classification

- Set $\mathcal{H} = \{x \mapsto h(x, y)\}$
- Set $\rho > 0$, then: whp, $\forall h$

$$L(h) \leq \widehat{L}_{n,\rho}(h) + \frac{2K^2}{\rho} R_n(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

Multiclass algorithms

- Intrinsic multiclass: SVM, boosting
- Aggregation of binary classifiers: one-versus-all, one-versus-one, error-correction codes

Chapter 4 - Advanced topics

Structured prediction

Setup

- Labels are sequences! text, DNA sequences, logs, ... or graphs!
- Idea: consider a joint feature vector $\psi(x, y) \in \mathbb{R}^p$
- Predictor class parameterized by a vector $w \in \mathbb{R}^p$

$$h(x) = \arg\max_y w^T \psi(x, y)$$

- SVM-type formulation of regularized learning

$$\min_w \left\{ \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \max_{y \neq y_i} \max\{0, 1 - w^T(\psi(x_i, y_i) - \psi(x_i, y))\} \right\}$$

Chapter 4 - Advanced topics

Sequential learning

Learning framework (1) - aggregation of weak predictors

- Consider $h_1, \ldots, h_M$ weak predictors (taking values in $[-1, 1]$) obtained by pretesting
- Denote by $H(x) = \big(h_1(x), \ldots, h_M(x)\big)^T$ the prediction vector, for any $x \in \mathbb{R}^d$
- For any $\lambda > 0$, define the λ-simplex of $\mathbb{R}^M$

$$\Theta = \Theta_{M,\lambda} = \left\{ \theta = (\theta^{(1)}, \ldots, \theta^{(M)})^T \in \mathbb{R}_+^M \ : \ \sum_{i=1}^{M} \theta^{(i)} = \lambda \right\}$$

- Search space:

$$\mathcal{F} = \mathcal{F}_{M,\lambda} = \left\{ f_\theta = \theta^T H \ : \ \theta \in \Theta \right\}$$

- Issue: find the optimal combination

Learning framework (2) - Sequential setup

- Want to optimize:

$$A(\theta) = \mathbb{E}\left(\varphi(-Y \cdot \theta^T H(X))\right)$$

- Observations are: independent and instant realizations of the subgradient of A

$$u_t(\theta) = \varphi'\left(-Y_t \theta^T H(X_t)\right) Y_t H(X_t) \in \mathbb{R}^M$$

where φ' is a monotone version of the derivative of φ

- Output: at every time t

$$g_t(x) = 2\mathbb{I}\{\widehat{\theta}_t H(x) > 0\} - 1$$

Notations

- Assume Θ is convex and closed in $(\mathbb{R}^M, \|\cdot\|_1)$
- Define the proxy function $V : \Theta \to \mathbb{R}_+$ convex and continuous
- Let $\beta > 0$, and W_β the β-convex conjugate of V: for any $z \in (\mathbb{R}^M, \|\cdot\|_\infty)$

$$W_\beta(z) = \sup_{\theta \in \Theta} \{-z^T\theta - \beta V(\theta)\}$$

Definition - Strong convexity

- Let $\alpha > 0$, set $V : \Theta \to \mathbb{R}_+$ to be convex
- The function V is said to be α-strongly convex relatively to $\|\cdot\|_1$ if: $\forall x, y \in \Theta$, $\forall s \in (0, 1]$,

$$V(sx + (1-s)y) \leq sV(x) + (1-s)V(y) - \frac{\alpha}{2}s(1-s)\|x-y\|_1^2$$

Key assumption

- We assume that $V : \Theta \to \mathbb{R}$ satisfies the following conditions:
  - V is C^1
  - V is α-strongly convex
  - V admits a unique minimum

Mirror Descent Algorithm (MDA)

- Parameters: (γ_i), (β_i)
- Initialization: $\theta_0 \in \Theta$, $\zeta_0 = 0 \in \mathbb{R}^M$
- For t=1, . . . , T:

$$\begin{aligned} \zeta_i &= \zeta_{i-1} + \gamma_i \\ \theta_i &= -\nabla W_{\beta_i}(\zeta_i) \end{aligned}$$

- Output : $\widehat{\theta}_T = \dfrac{\sum_{t=1}^{T} \gamma_t \theta_{t-1}}{\sum_{t=1}^{T} \gamma_t}$

Heuristics (1)

- Consider affine approximations of A at θ_{t-1} available at time $t \leq \tau - 1$:

$$\psi_t(\theta) = A(\theta_{t-1}) + (\theta - \theta_{t-1})^\tau \nabla A(\theta_{t-1})$$

and the convex average:

$$\overline{\psi}_\tau(\theta) = \frac{\sum_{t=1}^{\tau} \gamma_t \psi_t(\theta)}{\sum_{t=1}^{\tau} \gamma_t}$$

- Next "ideal" point:

$$"\theta_\tau" = \arg\min_{\theta \in \Theta} \left\{ \theta^T \sum_{t=1}^{\tau} \gamma_t \nabla A(\theta_{t-1}) + \beta_\tau V(\theta) \right\}$$

Heuristics (2)

- Since ∇A is not observed, it is replaced by $u_t(\theta)$ which leads to:

$$\theta_\tau = \arg\min_{\theta \in \Theta} \left\{ \theta^T \sum_{t=1}^{\tau} \gamma_t u_t(\theta_{t-1}) + \beta_\tau V(\theta) \right\}$$

which can be written as (after gradient step in MDA updates):

$$\theta_\tau = \arg\max_{\theta \in \Theta} \left\{ -\zeta_\tau^T \theta - \beta_\tau V(\theta) \right\}$$

Key property for convergence

- Under the assumptions on V, we have the following properties for W_β:
  - W_β is C^1
  - ∇W_β is $\frac{1}{\alpha\beta}$-Lipschitz:

$$\|\nabla W_\beta(z) - \nabla W_\beta(z')\|_1 \leq \frac{1}{\alpha\beta}\|z - z'\|_\infty$$

Main result

- Consider φ a loss function, $M \geq 2$ and $\lambda > 0$
- Then, for any $t \geq 1$, we have:

$$\mathbb{E}A(\widehat{\theta}_t) - \min_{\theta \in \Theta} A(\theta) \leq C\sqrt{\log(M)}\frac{\sqrt{t+1}}{t}$$

where $C = 2\lambda \sup_{[-\lambda,\lambda]} |\varphi'|$

Choice of the proxy function

- Entropic:

$$V(\theta) = \lambda \log\left(\frac{M}{\lambda}\right) + \sum_{j=1}^{M} \theta^{(j)} \log(\theta^{(j)})$$

- L_p with $p = 1 + \frac{1}{\log M}$

$$V(\theta) = \frac{1}{2\lambda^2} \|\theta\|_p^2$$

- L_p with $p = 1 + \frac{1}{\log M}$

$$V(\theta) = C_0 + C_1 \|\theta\|_p^p$$

where $C_0 = -\lambda^2 / es(s+1)$ and $C_1 = -\lambda^{1-s}/s(s+1)$, with $s = 1/\log(M)$

Conclusion of the course

- Only an **introduction** to Statistical Learning theory
- Essential background developed from 1995 to 2005 but still of intense activity on specific problems
- Purpose of the theory:
  - grasp the concepts behind the algorithms
  - but also important role in problem formulation (sampling/performance metrics/constraints)
- Data don't know what they are!
- ... and neither Machine learning knows what data are!

Join MLMDA team (about 20 people!) - Internship topics

- Far from the content of this course!
- Topic #1 - network science, contagion phenomena applied to epidemics, social networks, and transportation networks
- Topic #2 - representation and metric learning applied to functional data and applications to clinical assessment
- Topic #3 - reproducible research in Machine Learning
- Full range from advanced theory to interdisciplinary and highly collaborative projects, including spin-offs incubated in the lab
- New opportunities with the ATOS-CEA industrial chair starting in January 2017