Advanced Topics in Data Sciences

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 09: Overview of Learning Theory

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-731 (Spring 2016)

lions@epfl

MARIE CURIE ACTIONS

FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

erc

EPFL

Outline

This lecture:

1. The probably approximately correct (PAC) learning framework.
2. Empirical risk minimization (ERM).
3. Approximation and estimation errors.
4. Structural risk minimization (SRM).
5. Convex surrogate functions.
6. Stability and generalization.

Recommended reading materials

1. Chapters 2–4 in S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning*, Cambridge Univ. Press, 2014.

2. S. Boucheron *et al.*, "Theory of classification: A survey of some recent advances," *ESIAM: Probab. Stat.*, 2005.

The PAC Learning Framework

The standard statistical learning model

- **Training Data:** $\mathcal{D}_n := \{Z_i : 1 \leq i \leq n\} \sim$ i.i.d. unknown $\mathbb{P}$ on $\mathcal{Z}$
- **Hypothesis Class:** $\mathcal{H}$ a set of hypotheses h
- **Loss Function:** $f : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$
- **Risk:** $F(h) := \mathbb{E}_{\mathbb{P}} f(h, Z)$, where $Z \sim \mathbb{P}$ is independent of $\mathcal{D}_n$
- **Goal:** Find a "good hypothesis" $\hat{h}_n \in \mathcal{H}$ based on $\mathcal{D}_n$ such that $F(\hat{h}_n)$ is "small."

Observation

Statistical learning corresponds to solving the optimization problem

$$h^\star \in \arg\min_{h \in \mathcal{H}} F(h).$$

However, the optimization problem is not explicitly formulated, because $\mathbb{P}$ is unknown.

Example: Binary classification

- **Training Data:** $\mathcal{D}_n = \{Z_i = (X_i, Y_i) : 1 \leq i \leq n\}$
  - $X_i \in \mathbb{R}^p$ are images.
  - Each $Y_i \in \{0, 1\}$ labels whether there is a cat in the image X_i or not.
- **Hypothesis Class:** $\mathcal{H}$ a set of classifiers $h : \mathcal{X} \to \{0, 1\}$
  - $\mathcal{H}$ can be a set of linear classifiers, a reproducing kernel Hilbert space, or all possible realizations of a deep network.
- **Loss Function:** Binary loss $f(h, Z) := \mathbb{1}_{\{Y_i \neq h(X_i)\}}$, where $Z \sim \mathbb{P}$ is independent of $\mathcal{D}_n$
- **Risk:** $F(h) := \mathbb{E}_{\mathbb{P}} f(h, Z)$, which is the probability of false classification

Observation

The classifier that minimizes the risk is the **Bayes classifier**,

$$h(x) = \mathbb{1}_{\{\mathbb{P}(Y=1|X=x) \geq 1/2\}},$$

which is unfortunately intractable, since $\mathbb{P}$ is unknown.

Standard statistics approach: Logistic regression

1. Consider the class of linear classifiers $\mathcal{H} := \{\mathbb{1}_{\{\langle\cdot,\theta\rangle \leq 0\}} : \theta \in \Theta\}$ for some parameter space $\Theta \subseteq \mathbb{R}^p$.

2. **Assume** the statistical model (canonical generalized linear model [15])

$$\mathsf{P}\left(Y_i = 1 | X_i = x_i\right) = 1 - \mathsf{P}\left(Y_i = 0 | X_i = x_i\right) = \frac{1}{1 + \exp(-\langle x_i, \theta^\natural \rangle)},$$

   for some $\theta^\natural \in \Theta$. (See [9] for a Bayesian interpretation.)

3. Compute the **maximum-likelihood** estimator

$$\hat{\theta}_n \in \arg\min_{\theta \in \Theta} L_n(\theta),$$

   where L_n is the negative log-likelihood function.

4. Output the classifier $\hat{h}_n(\cdot) = \mathbb{1}_{\{\langle\cdot,\hat{\theta}_n\rangle \leq 0\}}$.

Question

Why should we assume this specific statistical model?

Possibly approximately correct (PAC) learnability

Definition (PAC learnability [20])

Assume that $Y_i = g(X_i)$ for some *deterministic function* g, and that $g \in \mathcal{H}$. (Hence zero risk is possible.)

A hypothesis class $\mathcal{H}$ is PAC learnable, if there exist an algorithm $\mathcal{A}_{\mathcal{H}} : \mathcal{Z}^n \to \mathcal{H}$ and a function $n_{\mathcal{H}}(\varepsilon, \delta)$, such that for **every probability distribution** $\mathbb{P}$ and every $\varepsilon, \delta \in (0, 1)$, if $n \geq n_{\mathcal{H}}(\varepsilon, \delta)$, we have

$$F(\mathcal{A}(\mathcal{D}_n)) \leq \varepsilon, \quad (\textbf{approximately correct})$$

with probability at least $1 - \delta$ (**probably**).

- The original definition also requires $\mathcal{A}_{\mathcal{H}}$ to be polynomial time, which we omit here for simplicity.
- The quantity $n_{\mathcal{H}}(\varepsilon, \delta)$ is called the **sample complexity**.

Questions

1. What if g is not contained in $\mathcal{H}$?
2. What if Y_i is a general random variable?

Agnotic PAC learnability

Definition (Agnostic PAC learnability [7])

A hypothesis class $\mathcal{H}$ is agnostic PAC learnable, if there exist an algorithm $\mathcal{A}_{\mathcal{H}} : \mathcal{Z}^n \to \mathcal{H}$ and a function $n_{\mathcal{H}}(\varepsilon, \delta)$, such that for every probability distribution $\mathbb{P}$ and every $\varepsilon, \delta \in (0, 1)$, if $n \geq n_{\mathcal{H}}(\varepsilon, \delta)$, we have

$$F(\mathcal{A}(\mathcal{D}_n)) - \inf_{h \in \mathcal{H}} F(h) \leq \varepsilon,$$

with probability at least $1 - \delta$.

A distribution-dependent and localized formulation [2, 3, 10, 11]

Given an algorithm $\mathcal{A}_{\mathcal{H}} : \mathcal{Z}^n \to \mathcal{H}$, show that for every probability distribution $\mathbb{P}$ and every $\delta \in (0, 1)$, we have

$$F(\mathcal{A}(\mathcal{D}_n)) - \inf_{h \in \mathcal{H}} F(h) \leq \varepsilon_n(\mathbb{P}, h^\star; \mathcal{H}, \delta) \to 0,$$

with probability at least $1 - \delta$, where $h^\star = \arg\min_{h \in \mathcal{H}} F(h)$ (assuming uniqueness).

- The quantity $F(\mathcal{A}(\mathcal{D}_n)) - \inf_{h \in \mathcal{H}} F(h)$ is called the **excess risk**.

Other examples

Linear regression

- Training data: $Z_i = (X_i, Y_i) \in \mathcal{Z} = \mathbb{R}^p \times \mathbb{R}$
- Hypothesis class: $\mathcal{H} = \{h_\theta(\cdot) = \langle \cdot, \theta \rangle : \theta \in \Theta\}$ for some $\Theta \subseteq \mathbb{R}^p$
- Loss function: Square error $f(h_\theta, z) = (y - \langle x, \theta \rangle)^2$

Density estimation

- Training data: $Z_i \in \mathbb{R}$
- Hypothesis class: A class of probability densities $\mathcal{P}$
- Loss function: Negative log-likelihood $f(p, z) = -\log p(z)$

K-means clustering/Vector quantization

- Training data: $Z_i \in \mathbb{R}^p$
- Hypothesis class: A class of subsets of $\mathbb{R}^p$ of cardinality K
- Loss function: $f(h, z) = \min_{c \in h} \|c - z\|_2^2$

Empirical Risk Minimization

Empirical risk minimization (ERM)

Recall that since $\mathbb{P}$ is assumed unknown, we cannot directly solve the risk minimization problem

$$h^\star \in \arg\min_{h \in \mathcal{H}} F(h) := \mathbb{E} f(h, Z).$$

However, we can consider the empirical risk minimization problem as an approximate,

$$\hat{h}_n \in \arg\min_{h \in \mathcal{H}} \hat{F}_n(h) := \frac{1}{n} \sum_{i \leq n} f(h, Z_i).$$

This is called the **ERM principle**, due to Vapnik and Chervonenkis.

Observation

By the strong law of large numbers (LLN), we know that $\hat{F}_n(h) \to F(h)$ almost surely for every $h \in \mathcal{H}$.

Question

Is the strong LLN argument enough to conclude that the ERM principle allows learnability?

Two notions of convergence

Definition (Convergence implied by the strong LLN)

For every $h \in \mathcal{H}$ and every probability distribution $\mathbb{P}$, there exists a function $n_{\mathcal{H}}(\varepsilon, \delta; h, \mathbb{P})$, such that for every $\varepsilon, \delta \in (0, 1)$, if $n \geq n_{\mathcal{H}}(\varepsilon, \delta; h, \mathbb{P})$, we have

$$|\hat{F}_n(h) - F(h)| \leq \varepsilon,$$

with probability at least $1 - \delta$.

Definition (Uniform convergence)

A hypothesis class $\mathcal{H}$ has the uniform convergence property, if there exists a function $n_{\mathcal{H}}(\varepsilon, \delta)$, such that for every $\varepsilon, \delta \in (0, 1)$ and any probability distribution $\mathbb{P}$, if $n \geq n_{\mathcal{H}}(\varepsilon, \delta)$, we have

$$\sup_{h \in \mathcal{H}} |\hat{F}_n(h) - F(h)| \leq \varepsilon,$$

with probability at least $1 - \delta$.

- Such an $\mathcal{H}$ with the uniform convergence property is called a **uniformly Glivenko-Cantelli class**.

Uniform convergence implies learnability

Proposition

For any $\varepsilon > 0$, if

$$\sup_{h \in \mathcal{H}} |\hat{F}_n(h) - F(h)| \leq \varepsilon,$$

then for any $h^\star \in \arg\min_{h \in \mathcal{H}} F(h)$, we have

$$F(\hat{h}_n) - F(h^\star) \leq 2\varepsilon.$$

Proof.

$$\begin{aligned} F(\hat{h}_n) - F(h^\star) &= F(\hat{h}_n) - \hat{F}_n(\hat{h}_n) + \hat{F}_n(\hat{h}_n) - \hat{F}_n(h^\star) + \hat{F}_n(h^\star) - F(h^\star) \\ &\leq 2 \sup_{h \in \mathcal{H}} |\hat{F}_n(h) - F(h)|. \end{aligned}$$

□

Observation

Uniform convergence property is sufficient for learnability.

⋆ Learnability, ERM, and uniform convergence

Theorem (See, e.g., [17])

Assume that the hypothesis class $\mathcal{H}$ consists of only $\{0, 1\}$-valued functions, and f is the $0 - 1$ loss. The following statements are equivalent.

1. *The hypothesis class is agnostic PAC learnable.*
2. *The ERM is a good PAC learner.*
3. *The hypothesis class has the uniform convergence property.*

Fact

Unfortunately, computing the corresponding ERM is in general NP-hard [8].

Fact

In general, uniform convergence may not be necessary for learnability [18].

Uniform convergence property of a finite bounded hypothesis class

Proposition

Assume that the hypothesis class $\mathcal{H}$ consists of a finite number of functions taking values in $[0,1]$. Then $\mathcal{H}$ satisfies the uniform convergence property with

$$n_{\mathcal{H}}(\varepsilon, \delta) = \frac{\log(2|\mathcal{H}|/\delta)}{2\varepsilon^2}.$$

The proposition is a simple consequence of Hoeffding's inequality and the union bound.

Theorem (Hoeffding's inequality (see, e.g., [14]))

Let $(\xi_i)_{1 \le i \le m}$ *be a sequence of independent* $[0,1]$*-valued random variables. Let* $S_n := (1/n) \sum_{1 \le i \le n} (\xi_i - \mathbb{E}\,\xi_i)$*. Then for any* $t > 0$, $\mathbb{P}\left(|S_n| \ge t\right) \le 2 \exp\left(-2nt^2\right)$.

Proof

Proof of the proposition.

Define $\xi_i(h) = f(h, x_i)$, and define $S_n(h) := (1/n) \sum_{1\leq i\leq n}(\xi_i(h) - \mathbb{E}\,\xi_i(h))$ for every $h \in \mathcal{H}$. Notice that then

$$\sup_{h\in\mathcal{H}} |S_n(h)| = \sup_{h\in\mathcal{H}} |\hat{F}_n(h) - F(h)|.$$

By the union bound and Hoeffding's inequality, we have for any $t > 0$,

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} |S_n(h)| \geq t\right) \leq \sum_{h\in\mathcal{H}} \mathbb{P}\left(|S_n(h)| \geq t\right) \leq |\mathcal{H}| \cdot 2\exp\left(-2nt^2\right).$$

Hence it suffices to choose

$$n_{\mathcal{H}}(\varepsilon, \delta) = \frac{\log(2|\mathcal{H}|/\delta)}{2\varepsilon^2}.$$

□

Necessity of choosing a not-too-big hypothesis class

We may write the proposition in another way:
For every probability distribution $\mathbb{P}$ and every $\delta \in (0,1)$, the ERM satisfies

$$\sup_{h \in \mathcal{H}} |\hat{F}_n(h) - F(h)| \leq \varepsilon_n := \sqrt{\frac{\log(2|\mathcal{H}|/\delta)}{2n}},$$

with probability at least $1-\delta$.

Observation

If $|\mathcal{H}|$ is large, we need a large number of training data of the order $O(\log|\mathcal{H}|)$ to achieve a small excess risk ε_n.

Otherwise, if ε_n is large, the values of $\hat{F}_n$ and F can be very different on certain hypotheses, and **overfitting** occurs.

Question

What if $\mathcal{H}$ is too small?

* What if $|\mathcal{H}|$ is not finite?

Consider the binary classification problem, in which $\mathcal{H}$ is a set of $\{0,1\}$-valued functions, and f is the $0-1$ loss.

Definition (Shattering coefficient)

The shattering coefficient of a hypothesis class $\mathcal{H}$ is defined as

$$S_n(\mathcal{H}) := \sup_{x_1,\ldots,x_n \in \mathcal{X}} |\{(h(x_i))_{1\leq i\leq n} : h \in \mathcal{H}\}|.$$

Definition (Vapnik-Chervonenkis (VC) dimension)

The VC dimension of a hypothesis class $\mathcal{H}$, denoted by $\mathrm{VC}(\mathcal{H})$, is defined as the largest integer k such that $S_k(\mathcal{H}) = 2^k$. If $S_k(\mathcal{H}) = 2^k$ for all k, then $\mathrm{VC}(\mathcal{H}) := \infty$.

Theorem ([21])

Let $\mathcal{H}$ be a hypothesis class with VC dimension d. Then

$$\sup_{h\in\mathcal{H}} |\hat{F}_n(h) - F(h)| \leq 2\sqrt{\frac{2d\log(2en/d)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}},$$

with probability at least $1-\delta$.

Model Selection and Structural Risk Minimization

Approximation and estimation errors

Let h_{opt} be a **global** minimizer of the risk $F(\cdot)$ which is not necessarily in $\mathcal{H}$.
Let $h^\star$ be a minimizer of the risk $F(\cdot)$ on $\mathcal{H}$.
Then we can write

$$F(\hat{h}_n) - F(h_{\text{opt}}) = F(\hat{h}_n) - F(h^\star) + F(h^\star) - F(h_{\text{opt}}).$$

Definition (Approximation error)

The approximation error is defined as $\mathcal{E}_{\text{app}} = F(h^\star) - F(h_{\text{opt}})$.

- The approximation error is fixed given a hypothesis class $\mathcal{H}$.
- The ERM can yield small risk only if $\mathcal{H}$ contains a "good enough" hypothesis.

Definition (Estimation error)

The estimation error is defined as $\mathcal{E}_{\text{est}} = F(\hat{h}_n) - F(h^\star)$.

- The estimation error decreases with the training data size.

Observation

If we shrink the hypothesis class, while the estimation error $\mathcal{E}_{\text{est}}$ can be smaller, doing so can only increase the approximation error $\mathcal{E}_{\text{app}}$.

Model selection

Model selection seeks a balance between approximation and estimation errors.

The model selection problem

Let $\mathcal{H}$ be a hypothesis class. Consider a countable family of sub-classes $\{\mathcal{H}_k : k \in \mathcal{K}\}$ such that $\bigcup_{k\in\mathcal{K}} \mathcal{H}_k = \mathcal{H}$. Denote by $\hat{h}_{n,k}$ an empirical risk minimizer chosen based on $\mathcal{D}_n$ in $\mathcal{H}_k$ for all $k \in \mathcal{K}$.

The model selection problem asks to choose a $\hat{k}_n \in \mathcal{K}$ based on $\mathcal{D}_n$, such that

$$F(\hat{h}_{n,\hat{k}_n}) - F(h^\star) \leq C \inf_{k\in\mathcal{K}} \left(\inf_{h\in\mathcal{H}_k} F(h) - F(h^\star) + \tilde{\pi}_n(k) \right),$$

with high probability for some constant $C > 0$ and $\tilde{\pi}_n(k) > 0$.

- Such an inequality on $F(\hat{h}_{n,\hat{k}_n}) - F(h^\star)$ is called an **oracle inequality**.
- If $C = 1$, the oracle inequality is called **sharp**.

Structural risk minimization (SRM)

The idea of structural risk minimization is to **minimize a risk estimate**.

Structural risk minimization (see, e.g., [22])

1. Choose
$$\hat{k}_n \in \arg\min_{k \in \mathcal{K}} (\hat{F}_n(\hat{h}_{n,k}) + \pi_n(k)),$$
where $\pi_n(k)$ is some good estimate of $F(\hat{h}_{n,k}) - \hat{F}_n(\hat{h}_{n,k})$.
2. Output $\hat{h}_n = \hat{h}_{n,\hat{k}_n}$.

- Computational complexity is completely ignored here.

Structural risk minimization (SRM) contd.

Theorem ([1])

Suppose there exists a double sequence $(R_{n,k})_{n\in\mathbb{N},k\in\mathcal{K}}$*, such that for every* $n \in \mathbb{N}$*,* $k \in \mathcal{K}$*, and* $\varepsilon > 0$*,*

$$\mathbb{P}\left(F(\hat{h}_{n,k}) > R_{n,k} + \varepsilon\right) \leq \alpha_n \exp(-2\beta_n \varepsilon^2),$$

for some constants $\alpha_n, \beta_n > 0$*. Set* $\pi_n(k) := R_{n,k} - \hat{F}_n(\hat{h}_{n,k}) + \sqrt{\beta_n^{-1} \log k}$*. Then we have*

$$F(\hat{h}_n) < \inf_k \left(\inf_{h\in\mathcal{H}_k} F(h) + \pi_n(k) + \sqrt{\frac{\log k}{n}} \right) + \varepsilon,$$

with probability at least $1 - 2\alpha_n \exp(-\beta_n \varepsilon^2/2) - 2\exp(-n\varepsilon^2/2)$*.*

Observation

- The risk bound based on the VC dimension may be used [13, 22], but it can be loose since the bound is for the worst case.
- Hence it is important to find sharp **data dependent** risk estimates. See [12] for some recent advances.

Convex Surrogate Functions

Logistic regression as a learning algorithm

Logistic regression as a learning algorithm

Given training data $(x_i, y_i) \in \mathbb{R}^p \times \{\pm 1\}$, $1 \leq i \leq n$.

Given a hypothesis class $\{\mathrm{sign}(\langle x, \theta \rangle) : \theta \in \Theta\}$ (linear classifiers) for some $\Theta \subset \mathbb{R}^p$.

Solve the **empirical risk minimization (?)** problem:

$$\hat{\theta}_n \in \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{1 \leq i \leq n} \log\left[1 + \exp\left(-y_i \langle x_i, \theta \rangle\right)\right].$$

Output the classifier $\hat{h}_n(x) = \mathrm{sign}(\langle x, \hat{\theta}_n \rangle)$.

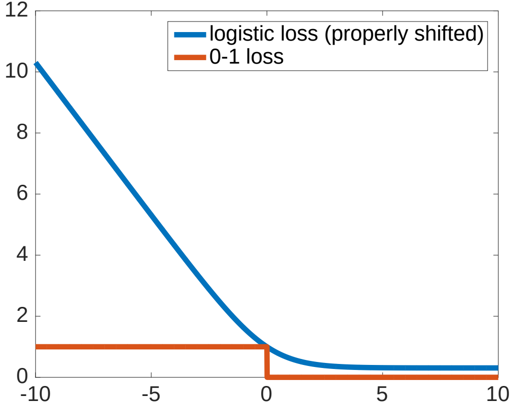
Observation

Unlike the empirical risk minimization problem with the $0-1$ loss, the logistic regression approach yields a **convex optimization problem** that can be efficiently solved (when Θ is also convex).

Question

Why does logistic regression work for binary classification?

Intuition



- Logistic loss: $\phi(t) = \log(1 + \exp(-t))$
- $0 - 1$ loss: $\phi(t) = \mathbb{1}_{\{t \leq 0\}}$

  (For logistic regression, t corresponds to $y\langle x, \theta \rangle$.)

Soft classification

Consider a **cost function** $g : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$.

Definition (Margin-based cost [6])

A cost function g is margin-based, if it can be written as $g(h, z) = \phi(yh(x))$ for some function ϕ.

Example

In logistic regression, $\phi(t) = \log(1 + \exp(-t))$, and $h \in \mathcal{H} = \{\langle \cdot, \theta \rangle : \theta \in \Theta\}$.

The corresponding **empirical cost minimization** problem is given by

$$\hat{h}_n \in \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{1 \leq i \leq n} \phi(y_i h(x_i)).$$

The corresponding **soft classifier** is given by

$$\tilde{h}_n(x) = \mathrm{sign}(\hat{h}_n(x)).$$

Zhang's lemma

Define

$$H_\phi(\eta,\alpha) = \eta\phi(\alpha) + (1-\eta)\phi(-\alpha), \quad \alpha^\star_\phi(\eta) = \arg\min_\alpha H_\phi(\eta,\alpha).$$

Let $F(h) = \mathbb{P}(\mathrm{sign}(h(X)) \neq Y)$ denote the risk function, and $G(h) = \mathbb{E}\,\phi(Yh(X))$ be the expected cost function.

Zhang's lemma ([23])

Assume that ϕ is convex, and $\alpha^\star_\phi(\eta) > 0$ when $\eta > 1/2$. If there exist $c > 0$ and $s \geq 1$ such that for all $\eta \in [0,1]$,

$$|1/2-\eta|^s \leq c^s \left[H(\eta,0) - H(\eta,\alpha^\star(\eta))\right]^{1/s},$$

Then for any hypothesis h,

$$F(h) - \min_h F(h) \leq 2c\left[G(h) - \min_h G(h)\right]^{1/s}.$$

Risk bound for ℓ_1-regularized logistic regression

- For the logistic regression, $c = 1/\sqrt{2}$ and $s = 2$.

Theorem

Consider the ℓ_1-regularized logistic regression with $\Theta = \{\theta : \|\theta\|_1 \leq \nu\}$ for some $\nu > 0$. Assume that $\|x\|_\infty \leq 1$ for all $x \in \mathcal{X}$. Then there exists a constant $C > 0$ depending only on p, such that with probability at least $1 - \delta$,

$$F(\hat{h}_n) - \inf_h F(h) \leq \; 4\left(\nu\sqrt{\frac{C}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}}\right)^{1/2} + \sqrt{2}\left[\left(\inf_{h \in \mathcal{H}} G(h)\right) - \left(\inf_h G(h)\right)\right]^{1/2}.$$

Proof.

Similar to Theorem 4.4 in [4]. □

- The right-hand side may be viewed as the sum of the estimation error and approximation error (w.r.t. the cost).

Other examples

Recall the risk bound

$$F(h) - \min_h F(h) \leq 2c \left[G(h) - \min_h G(h) \right]^{1/s}.$$

AdaBoost (See, e.g., [16])

Adaboost is equivalent to solving an empirical cost minimization problem with $\phi(t) = \exp(-t)$, for which $s = 2$ and $c = 1/\sqrt{2}$.

- Notice that in practice AdaBoost may not be implemented by directly solving the empirical cost minimization problem.

Support vector machine (See, e.g., [19])

The hinge cost function used by the support vector machine (SVM) corresponds to $\phi(t) = \max(0, 1-t)$, for which $s = 1$ and $c = 1/2$.

Stability

Analysis of SVM

A linear SVM is given by

$$\hat{\theta}_n \in \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{1 \leq i \leq n} \phi(y_i \langle x_i, \theta \rangle) + \lambda \|\theta\|_2^2 \,,$$

for some $\lambda > 0$, where $\phi(t) = \max(0, 1-t)$ is the hinge loss.

The output classifier is given by $\tilde{h}_n(\cdot) = \mathrm{sign}(\langle \cdot, \hat{\theta}_n \rangle)$.

Question

How do we analyze SVM, which is not exactly empirical cost minimization?

Idea

Instead of considering a class of algorithms, we may do an algorithm-wise analysis.

Stability implies generalization

Definition (Classification stability [5])

Consider the soft classification setting, where $\mathcal{H}$ is a class of soft classifiers (i.e., $\tilde{h}_n = \mathrm{sign}(\hat{h}_n)$). For any $\mathcal{D}_n = \{z_1, \ldots, z_n\}$, define $\mathcal{D}_n^{\setminus i}$ as $\mathcal{D}_n$ with the i-th element z_i removed. An algorithm $\mathcal{A}$ has classification stability with parameter $\beta > 0$, if for all $\mathcal{D}_n \subset \mathcal{Z}$ and for all $1 \leq i \leq n$,

$$\|\mathcal{A}(\mathcal{D}_n) - \mathcal{A}(\mathcal{D}_n^{\setminus i})\|_{L_\infty} \leq \beta.$$

Observation

Then by the triangle inequality,

$$\|\mathcal{A}(\mathcal{D}_n) - \mathcal{A}(\mathcal{D}_n \cup \{z\})\|_{L_\infty} \leq 2\beta, \quad \text{for all } z \in \mathcal{Z},$$

meaning the algorithm is robust to a small change of the training data.

Stability implies generalization contd.

Consider the $0-1$ loss $f(h,z) = \mathbb{1}_{\{\mathrm{sign}(h(x)) \neq y\}}$. Then the risk $F(h) = \mathbb{E} f(h,Z)$ is the probability of classification error.

Define the margin-based loss

$$f^{\gamma}(h,z) = \begin{cases} 1 & \text{for } yh(x) \neq 0 \\ 1 - yh(x)/\gamma & \text{for } 0 \leq yh(x) \leq \gamma \\ 0 & \text{for } yh(x) \geq \gamma \end{cases},$$

and the corresponding margin-based empirical risk $\hat{F}_n^{\gamma}(h) = (1/n)\sum_{1 \leq i \leq n} f_{\gamma}(h, z_i)$.

Theorem ([5])

Let $\mathcal{A}$ be a soft classification algorithm that possesses classification stability with parameter $\beta_n > 0$. Then for any $\gamma > 0$, $n \in \mathbb{N}$, and any $\delta \in (0,1)$,

$$F(\mathcal{A}(\mathcal{D}_n)) \leq \hat{F}_n^{\gamma}(\mathcal{A}(\mathcal{D}_n)) + 2\frac{\beta_n}{\gamma} + \left(1 + 4n\frac{\beta_n}{\gamma}\right)\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Observation

Notice that the uniform convergence property is not required.

Risk bound for the linear SVM

Recall that the linear SVM defines the algorithm $\mathcal{A}_{\mathsf{SVM}}(\mathcal{D}_n) = \langle \cdot, \hat{\theta}_n \rangle$, where

$$\hat{\theta}_n \in \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{1 \leq i \leq n} \phi(y_i \langle x_i, \theta \rangle) + \lambda \|\theta\|_2^2 .$$

Theorem

Assume that $\|x\|_2 \leq \kappa$ for all $x \in \mathcal{X}$ for some $\kappa > 0$. Then $\mathcal{A}_{\mathsf{SVM}}$ has classification stability with parameter $\beta_n = \kappa^2/(2\lambda n)$. Hence for any $n \in \mathbb{N}$ and $\delta \in (0,1)$,

$$F(\mathcal{A}_{\mathsf{SVM}}(\mathcal{D}_n)) \leq \hat{F}_n^1(\mathcal{A}_{\mathsf{SVM}}(\mathcal{D}_n)) + \frac{\kappa^2}{\lambda n} + \left(1 + \frac{2\kappa^2}{\lambda}\right) \sqrt{\frac{\log(1/\delta)}{2n}},$$

with probability at least $1 - \delta$.

Proof.

Similar to Example 2 in [5]. □

Review

Review

- Learning theory is concerned with developing learning algorithms that have **distribution-free** guarantees.
- The **ERM** principle provides a principled approach, if the **uniform convergence** property holds.
- **SRM** is an extension of the ERM principle that seeks a balance between the **estimation error** and the **approximation error**.
- In practice, we may replace the loss by an **convex surrogate** to yield an efficiently solvable **empirical cost minimization** problem.
- **Stability** provides another algorithm-wise analysis framework.

References I

[1] Peter L. Bartlett, Stéphane Boucheron, and Gábor Lugosi.
Model selection and error estimation.
Mach. Learn., 48:85–113, 2002.

[2] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson.
Local Rademacher complexities.
Ann. Stat., 33(4):1497–1537, 2005.

[3] Peter L. Bartlett and Shahar Mendelson.
Rademacher and Gaussian complexities: Risk bounds and structural results.
J. Mach. Learn. Res., 3, 2002.

[4] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi.
Theory of classification: A survey of some recent advances.
ESAIM: Probab. Stat., 9:323–375, November 2005.

[5] Olivier Bousquet and André Elisseeff.
Stability and generalization.
J. Mach. Learn. Res., 2:499–526, 2002.

[6] Corinna Cortes and Vladimir Vapnik.
Support-vector networks.
Mach. Learn., 20:273–297, 1995.

References II

[7] David Haussler.
Decision theoretic generalizations of the PAC model for neural net and other learning applications.
Inf. Comput., 100:78–150, 1992.

[8] Klaus-U. Höffgen and Hans-U. Simon.
Robust trainability of single neurons.
J. Comput. Syst. Sci., 50:114–125, 1995.

[9] Michael I. Jordan.
Why the logistic function? a tutorial discussion on probabilities and neural networks.
MIT Computational Cognitive Science report 9503, 1995.

[10] V. Koltchinskii and D. Panchenko.
Rademacher processes and bounding the risk of function learning.
2004.
arXiv:math/0405338v1 [math.PR].

[11] Vladimir Koltchinskii.
Local Rademacher complexities and oracle inequalities in risk minimization.
Ann. Stat., 34(6):2593–2656, 2006.

References III

[12] Vladimir Koltchinskii.
Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems.
Springer-Verl., Berlin, 2011.

[13] Gábor Lugosi and Kenneth Zeger.
Concept learning using complexity regularization.
IEEE Trans. Inf. Theory, 42(1):48–54, 1996.

[14] Pascal Massart.
Concentration Inequalities and Model Selection.
Springer-Verl., Berlin, 2007.

[15] P. McCullagh and J. A. Nelder.
Generalized Linear Models.
Chapman and Hall, London, second edition, 1989.

[16] Robert E. Schapire and Yoav Freund.
Boosting.
MIT Press, Cambridge, MA, 2012.

[17] Shai Shalev-Shwartz and Shai Ben-David.
Understanding Machine Learning.
Cambridge Univ. Press, Cambridge, UK, 2014.

References IV

[18] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan.
Learnability, stability and uniform convergence.
J. Mach. Learn. Res., 11:2635–2670, 2010.

[19] Ingo Steinwart and Andreas Christmann.
Support vector machines.
Springer, New York, NY, 2008.

[20] L. G. Valiant.
A theory of the learnable.
Commun. ACM, 27(11):1134–1142, November 1984.

[21] V. N. Vapnik and A. Ya. Chervonenkis.
On the uniform convergence of relative frequencies of events to their probabilities.
Theory Probab. Appl., XVI(2):264–280, 1971.

[22] Vladimir N. Vapnik.
Statistical Learning Theory.
John Wiley & Sons, New York, NY, 1998.

[23] Tong Zhang.
Statistical behavior and consistency of classification methods based on convex risk minimization.
Ann. Stat., 32(1):56–134, 2004.