Statistical Machine Learning

Christoph Lampert

I S T AUSTRIA

Institute of Science and Technology

Spring Semester 2015/2016 // Lecture 7

Overview (tentative)

Date		no.	Topic
Mar 01	Tue	1	A Hands-On Introduction
Mar 03	Thu	2	Bayesian Decision Theory
			Generative Probabilistic Models
Mar 08	Tue	3	Discriminative Probabilistic Models
			Maximum Margin Classifiers
Mar 10	Thu	4	Optimization, Kernel Classifiers
Mar 15	Tue	5	More Optimization; Model Selection
Mar 17	Thu	6	Beyond Binary Classification
Mar 21 – Apr 01			Spring Break
Apr 05	Tue	7	Learning Theory I
Apr 07	Thu	8	Learning Theory II
Apr 12	Tue	9	...overflow buffer...
Apr 14	Thu	10	Probabilistic Graphical Models
Apr 19	Tue	11	Deep Learning
Apr 21	Thu	12	Unsupervised Learning
until May 01			final project

What problems are "learnable"?

PAC Learning Scenario

- $\mathcal{X}$: input set, $\mathcal{Y}$: label set, **here**: $\mathcal{Y} = \{-1, 1\}$ or $\mathcal{Y} = \{0, 1\}$
- $p(x, y)$: data distribution (unknown to us)
- **for now: deterministic labels**, $y = f(x)$ for unknown $f : \mathcal{X} \to \mathcal{Y}$
- $\mathcal{D}_m = \{(x_1, y_1), \ldots, (x_m, y_m)\} \overset{i.i.d.}{\sim} p(x, y)$: training set
- $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$: loss function. **here:** $\ell(y, y') = [\![y \neq y']\!]$
- $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$: hypothesis set (the lerner's choice)
  e.g. "all linear classifiers in $\mathbb{R}^d$", or "all binary decision trees", ...

Quantity of interest:

- $\mathcal{R}_p(h) \;=\; \mathbb{E}_{(x,y)\sim p(x,y)}\, \ell(\, y, h(x)\,) \;=\; \Pr_{x\sim p(x)}\{\, f(x) \neq h(x)\,\}$

What does "learning" mean?

- We know: there is (at least one) $f : \mathcal{X} \to \mathcal{Y}$ that has $\mathcal{R}(f) = 0$.
- Can we find such f from $\mathcal{D}_m$? If yes, how large must m be?

Definition (Probably Approximately Correct (PAC) Learnability)

A hypothesis class $\mathcal{H}$ is called **PAC learnable** by an algorithm A, if

- for every $\epsilon > 0$ (accuracy → "approximate correct")
- and every $\delta > 0$ (confidence → "probably")

there exists an

- $m_0 = m_0(\epsilon, \delta) \in \mathbb{N}$ (minimal training set size)

such that

- for any probability distribution p over $\mathcal{X}$, and
- for any labeling function $f \in \mathcal{H}$, with $\mathcal{R}_p(f) = 0$,

when we run the learning algorithm A on a training set consisting of $m \geq m_0$ examples sampled i.i.d. from p, the algorithm returns a hypothesis $h \in \mathcal{H}$ that, with probability at least $1 - \delta$, fulfills $\mathcal{R}_p(h) \leq \epsilon$.

$$\forall m \geq m_0(\epsilon, \delta) \quad \Pr_{\mathcal{D}_m \sim p}[\,\mathcal{R}_d(A[\mathcal{D}_m]) > \epsilon\,] \leq \delta.$$

Note: for "efficient learning", A must run in poly$(m, \frac{1}{\epsilon}, \frac{1}{\delta},$ *"size of $\mathcal{D}_m$"*$)$

Empirical Risk Minimization

What *learning algorithm*?

Definition (Empirical Risk Minimization (ERM) Algorithm)

input hypothesis set $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$ (not necessarily finite)

input training set $\mathcal{D} = \{(x_1, y_1), \dots, (x_m, y_m)\}$

output $h \in \operatorname*{\mathbf{argmin}}_{h \in H} \frac{1}{m} \sum_{i=1}^{m} \ell(\, y_i, h(x_i)\,)$ (lowest training error)

ERM learns a classifier that has minimal training error.

- There might be multiple, we can't control which one.
- We saw already: ERM might well or might not work.
- Can we characterize when ERM works and when it fails?

Examples

A constant decision is PAC-learnable

- $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{\pm 1\}$, $\ell(y, y') = [\![y, y']\!]$
- $\mathcal{H} = \{h_+, h_-\}$ with $h_+(x) = +1$ and $h_-(x) = -1$
- p arbitrary

ERM needs only 1 example, then its solution is unique and perfect.

A parity bit is learnable

- $\mathcal{X} = \{0, 1\}^d$, $\mathcal{Y} = \{\pm 1\}$, $\ell(y, y') = [\![y, y']\!]$
- $\mathcal{H} = \{h_e, h_o\}$ with $h_e(x) = \otimes_{i=1}^d x_i$ and $h_o(x) = 1 - \otimes_{i=1}^d x_i$
- p arbitrary
- $\mathcal{D}_m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$

ERM needs only 1 example, then it's solution is unique and perfect.

Examples

Coordinate classifiers

- $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{\pm 1\}$, $\ell(y, y') = [\![y, y']\!]$
- $\mathcal{H} = \{h_1, \dots, h_d\}$ with $h_i(x) = \operatorname{sign} x[i]$

Lemma

If p is uniform in $[-1, 1]^d$, ERM works for $m_0(\epsilon, \delta) = \lceil \log_2 \frac{d-1}{\delta} \rceil$

Proof: blackboard/notes

Here: for general distributions, we might have to return hypothesis with $\epsilon > 0$, and m_0 will depend on ϵ.

Can we prove general statements?

Theorem (PAC Learnability of finite hypothesis classes)

Let $\mathcal{H} = \{h_1, \dots, h_K\}$ be a finite hypothesis class and $f \in \mathcal{H}$ (i.e. the true labeling function is one of the hypotheses).

Then $\mathcal{H}$ is PAC-learnable by the empirical risk minimization algorithm with $m_0(\epsilon, \delta) = \lceil \frac{1}{\epsilon}(\log(|\mathcal{H}| + \log(1/\delta)) \rceil$

Proof: blackboard/notes

Examples: Finite hypothesis classes

Model selection:

- Clients offer me trained classifiers: 1) *decision tree*, 2) *LogReg* or an 3) *SVM*? Which of the three should I buy?

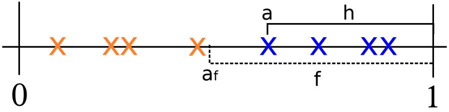
Finite precision:

- For $\mathcal{X} \subset \mathbb{R}^d$, the hypothesis set $\mathcal{H} = \{f(x) = \mathrm{sign}\langle w, x\rangle\}$ is infinite.
- But: on a computer, w is restricted to 64-bit `doubles`: $|\mathcal{H}_c| = 2^{64d}$.
  $m_0(\epsilon, \delta) = \frac{1}{\epsilon}(\log(|\mathcal{H}| + \log(1/\delta)) \approx \frac{1}{\epsilon}(44d + \log(1/\delta))$

Implementation:

- $\mathcal{H} = \{$ all algorithms implementable in 1 MB C-code $\}$ is finite.

Logarithmic dependence on $|\mathcal{H}|$ makes even large (finite) hypothesis sets (kind of) practical.

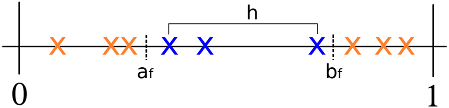


- $\mathcal{X} = [0,1]$, $\mathcal{Y} = \{0,1\}$,
- $\mathcal{H} = \{h_a(x) = [\![\, x \geq a \,]\!],$ for $0 \leq a \leq 1\}$,
- $f(x) = h_{a_f}(x)$ for some $0 \leq a_f \leq 1$.
- ERM rule: $\quad h = \operatorname*{\mathbf{argmin}}_{h_a \in H} \frac{1}{m} \sum_{i=1}^{m} [\![h_a(x_i) \neq y_i]\!]$,

  pick *smallest possible "$+1$" region* when not unique
  (to make algorithm deterministic): $a = \mathbf{min}_{\{i: y_i = 1\}} \{x_i\}$

Claim: ERM learns f (in the PAC sense). Proof: textbook...

Example: Learning Intervals

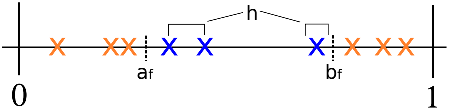


- $\mathcal{X} = [0,1]$, $\mathcal{Y} = \{0,1\}$,
- $\mathcal{H} = \Big\{ h_{[a,b]}(x) = [\![x \geq a \wedge x \leq b]\!],$ for $0 \leq a \leq b \leq 1 \Big\}$,
- $f(x) = h_{[a_f,b_f]}(x)$ for some $0 \leq a_f \leq b_f \leq 1$.
- training set $S = \{(x_1,y_1),\dots,(x_m,y_m)\}$
- ERM rule: $h = \operatorname*{\mathbf{argmin}}_{[a,b]} \frac{1}{m}\sum_{i=1}^{m} [\![h_{[a,b]}(x_i) \neq y_i]\!]$,

  pick *smallest possible "+1" interval* when not unique:
  $a = \mathbf{min}_{\{i:y_i=1\}}\{x_i\}$, $b = \mathbf{max}_{\{i:y_i=1\}}\{x_i\}$

Claim: ERM learns f in the PAC sense. Proof: textbook...

Example: Learning Unions of Intervals



- $\mathcal{X} = [0,1]$, $\mathcal{Y} = \{0,1\}$,
- $\mathcal{H} = \Big\{ h_{[a,b]}(x) \text{ for } \mathcal{I} = \{I_1,\dots,I_K\} \text{ for some } K \in \mathbb{N} \Big\}$,
  for $h_{\mathcal{I}}(x) = [\![\, x \in \bigcup_{k=1}^{K} I_k]\!]$ with $I_i = [a_k, b_k]$
- $f(x) = h_{\mathcal{I}_f}(x)$ for some set of intervals $\mathcal{I}_f$
- training set $S = \{(x_1,y_1),\dots,(x_m,y_m)\}$
- ERM rule: $$h = \operatorname*{\mathbf{argmin}}_{\mathcal{I}} \frac{1}{m}\sum_{i=1}^{m} [\![h_{\mathcal{I}}(x_i) \neq y_i]\!],$$

  pick *smallest possible "$+1$" region* when not unique

Claim: ERM fails to learn f in the PAC sense.
Proof: textbook... (but obvious: $h_{\text{ERM}} \equiv 0$ except in $x_1,\dots,x_m$)

There's No Free Lunch

Observation: ERM can learn all finite classes, but not some infinite ones.

Is there a better algorithm than ERM, one that *always works*?

There's No Free Lunch

Observation: ERM can learn all finite classes, but not some infinite ones.

Is there a better algorithm than ERM, one that *always works*?

No-Free-Lunch Theorem

- $\mathcal{X}$ input set, $\mathcal{Y} = \{0, 1\}$ label set, $\ell : \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}$: 0/1-loss,
- A an arbitrary learning algorithm for binary classification,
- m (training size) any number smaller than $|\mathcal{X}|/2$

There exists

- a data distribution p over $\mathcal{X} \times \mathcal{Y}$, and
- a function $f : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$ with $\mathcal{R}_p(f) = 0$, but

$$\Pr_{S \sim p^{\otimes m}} [\ \mathcal{R}_p(A[S]) \geq 1/8\ ] \geq 1/7.$$

Summary: For every learner, there exists a task on which it fails!

Agnostic PAC Learning

More realistic scenario: labeling isn't a deterministic function

- $\mathcal{X}$: input set
- $\mathcal{Y}$: output/label set, for now: $\mathcal{Y} = \{-1, 1\}$ or $\mathcal{Y} = \{0, 1\}$
- $p(x, y)$: data distribution (unknown to us)
- ~~**deterministic** labels, $y = f(x)$ for unknown $f : \mathcal{X} \to \mathcal{Y}$~~
- $S = \{(x_1, y_1), \dots, (x_m, y_m)\} \overset{i.i.d.}{\sim} p(x, y)$: training set
- $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$: loss function, $\ell(y, y') = [\![y \neq y']\!]$
- $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$: hypothesis set (the lerner's choice)

Quantity of interest:

- $\mathcal{R}_p(h) = \mathop{\mathbb{E}}_{(x,y)\sim p(x,y)} \ell(y, h(x)) = \Pr_{(x,y)\sim p(x,y)} \{h(x) \neq y\}$

What can we learn?

- there might not be any $f : \mathcal{X} \to \mathcal{Y}$ that has $\mathcal{R}(f) = 0$.
- but can we at least find the best h from the hypothesis set?

Definition (Agnostic PAC Learning)

A hypothesis class $\mathcal{H}$ is called **agnostic PAC learnable** by A, if

- for every $\epsilon > 0$ (accuracy → "approximate correct")
- and every $\delta > 0$ (confidence → "probably")

there exists an

- $m_0 = m_0(\epsilon, \delta) \in \mathbb{N}$ (minimal training set size)

such that

- for every probability distribution $p(x, y)$ over $\mathcal{X} \times \mathcal{Y}$,

when we run the learning algorithm A on a training set consisting of $m \geq m_0$ examples sampled i.i.d. from d, the algorithm returns a hypothesis $h \in \mathcal{H}$ that, with probability at least $1 - \delta$, fulfills

$$\mathcal{R}_p(h) \leq \min_{\bar{h} \in \mathcal{H}} \mathcal{R}_p(\bar{h}) + \epsilon.$$

$$\forall m \geq m_0(\epsilon, \delta) \quad \Pr_{S \sim p^{\otimes m}} \left[\mathcal{R}_p(A[S]) - \min_{\bar{h} \in \mathcal{H}} \mathcal{R}_p(\bar{h}) \; > \; \epsilon \right] \; \leq \; \delta.$$

Theorem (PAC Learnability of finite hypothesis classes)

Let $\mathcal{H} = \{h_1, \ldots, h_K\}$ be a finite hypothesis class.

Then $\mathcal{H}$ is agnostic PAC-learnable by ERM with $m_0(\epsilon, \delta) = \lceil \frac{2}{\epsilon^2}(\log(|\mathcal{H}| + \log(2/\delta)) \rceil$.

Proof sketch. Step 1: we bound $\mathcal{R}(h) - \hat{\mathcal{R}}_m(h)$ uniformly in h:

Lemma

For any $\epsilon > 0$, $\delta > 0$, the following inequality hold uniformly in $h \in \mathcal{H}$ with probability at least $1 - \delta$ w.r.t. $\mathcal{D}_m$:

$$|\mathcal{R}_p(h) - \hat{\mathcal{R}}_m(h)| \leq \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}}$$

Proof: blackboard/notes

Step 2: we use the lemma to bound the difference between

- $h_{\mathsf{ERM}} \in \mathbf{argmin}_{\bar{h} \in \mathcal{H}} \hat{\mathcal{R}}_m(\bar{h})$ (result of ERM)
- $h^* \in \mathbf{argmin}_{\bar{h} \in \mathcal{H}} \mathcal{R}_p(\bar{h})$ (if exists, otherwise use argument of arbitrarily close approximation)

$$\mathcal{R}_p(h_{\mathsf{ERM}}) - \mathcal{R}_p(h^*) \leq 2\sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}} \quad \overset{m \geq m_0}{\leq} \quad \epsilon$$

Vapnik-Chervonenkis (VC) dimension

Definition

Let $\mathcal{H} \subseteq \{\, \mathcal{X} \to \{0,1\} \,\}$ be a hypothesis class and $C = \{x_1, \ldots, x_m\} \subseteq \mathcal{X}$ be a finite set. The restriction of $\mathcal{H}$ to C is

$$\mathcal{H}_C = \left\{ \, \Big(h(x_1), h(x_2), \ldots, h(x_m) \Big) \; : \; h \in \mathcal{H} \right\} \subseteq \{0,1\}^m$$

Definition (Shattering)

A hypothesis class $\mathcal{H}$ **shatters** a finite set $C \subseteq \mathcal{X}$, if the restriction of $\mathcal{H}$ to C is the set of all possible labeling of C by $\{0,1\}$, i.e. $|\mathcal{H}_C| = 2^{|C|}$.

Definition (VC Dimension)

The **VC dimension** of a hypothesis class $\mathcal{H}$, denoted $\mathrm{VCdim}(\mathcal{H})$, is the maximal size of a set $C \subseteq \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size we say that $\mathrm{VCdim}(\mathcal{H}) = \infty$.

Lemma

For any finite $\mathcal{H}$, we have $\quad VCdim(\mathcal{H}) \leq \log_2 |\mathcal{H}|$.

Proof. $|\mathcal{H}_C| \leq |\mathcal{H}|$. So $|\mathcal{H}_C| = 2^{|C|}$ implies $|C| \leq \log_2 \mathcal{H}$

Lemma

Let $\mathcal{H} = \{h(x) = \text{sign}\langle w, x\rangle \ : w \in \mathbb{R}^d\}$ *be set of all linear classifiers in* $\mathbb{R}^d$. *Then* $VCdim(\mathcal{H}) = d$.

Proof. textbook...

Lemma

$\mathcal{X} = \mathbb{R}, \quad \mathcal{H} = \{h_\omega(x) = \text{sign}[\sin(\omega x)] : \omega \in \mathbb{R}\}. \quad VCdim(\mathcal{H}) = \infty$.

Proof. pick $C = \{1, \ldots, m\}$ and show that for each $(y_1, \ldots, y_m) \in \{\pm 1\}^m$ an ω exists such that $h_\omega(i) = y_i$.

Theorem (Fundamental Theorem of Statistical Learning (Subset))

Let $\mathcal{H} \subseteq \{ \mathcal{X} \to \{0,1\} \}$ be a hypothesis set, and let ℓ be the $0/1$-loss. Then, the following statements are equivalent:

- *$\mathcal{H}$ is PAC learnable.*
- *$\mathcal{H}$ is agnostic PAC learnable.*
- *Any ERM rule learns $\mathcal{H}$ in the PAC learning sense.*
- *Any ERM rule learns $\mathcal{H}$ in the agnostic PAC learning sense.*
- *$\mathcal{H}$ has finite VC-dimension.*

Proof. textbook...