Review of chapter2 $\sim$ chapter6

July 10, 2019

.

Chapter2 $\sim$ Chapter3

1. Data features:
   - Domain set: set of objects $\mathcal{X}$
   - Label set: set of objects $\mathcal{Y}$
   - Unknown distribution:
     - $\mathcal{D} \sim \mathcal{X}$, with unknown fixed mapping function $f : \mathcal{X} \to \mathcal{Y}$
     - $\mathcal{D} \sim (\mathcal{X}, \mathcal{Y})$
   - Training data: finite sequence
     $S = \mathcal{X} \times \mathcal{Y} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$
2. Learning framework:
   - Label function: $h : \mathcal{X} \to \mathcal{Y}$
   - Hypothesis set (or Hypothesis class): $\mathcal{H} = \{h_1, h_2, \ldots\}$
   - True error:
     $L_{\mathcal{D}}(h) := \underset{(x,y)\sim\mathcal{D}}{\mathbb{P}}[h(x) \neq y] := \mathcal{D}(\{(x, y) : h(x) \neq y\})$
   - Empirical error: $L_s(h) := \frac{|\{i\in[m]:h(x_i)\neq y_i\}|}{m}$, $[m] = \{1, \ldots, m\}$
   - Empirical risk minimization: $h_S^{ERM} = \arg\min_{h\in\mathcal{H}} L_S(h)$
   - ERM algorithm: $A^{ERM} : S \to h_S^{ERM}$

Chapter2 $\sim$ Chapter3

Definition 3.1 (PAC Learnability). A hypothesis class $\mathcal{H}$ is PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$, for every distribution $\mathcal{D}$ over $\mathcal{X}$, and for every labeling function $f : \mathcal{X} \to \{0,1\}$, if the realizable assumption holds with respect to $\mathcal{H}, \mathcal{D}, f$, then when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$ and labeled by f, the algorithm returns a hypothesis h such that, with probability of at least $1-\delta$ (over the choice of the examples), $L_{(\mathcal{D},f)}(h) \leq \epsilon$.

A hypothesis class $\mathcal{H}$ is *PAC learnable* if:
$\exists m_{\mathcal{H}}(\delta, \epsilon) \to \mathbb{N}$ satisfies:
$\forall \epsilon, \delta \in (0,1), \{S : |S| \geq m_{\mathcal{H}}(\epsilon, \delta), S \sim \mathcal{D}^m\}$,
$\mathbb{P}\{\exists h_S = A(S) \in \mathcal{H}, L_{(\mathcal{D},f)}(h_S) \leq \epsilon\} \geq 1 - \delta$

Chapter2 $\sim$ Chapter3

Definition 3.3 (Agnostic PAC Learnability). A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns a hypothesis h such that, with probability of at least $1 - \delta$ (over the choice of the m training examples),

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon.$$

A hypothesis class $\mathcal{H}$ is *agnostic PAC learnable* if:
$\exists m_{\mathcal{H}}(\delta, \epsilon) \to \mathbb{N}$ satisfies:
$\forall \epsilon, \delta \in (0,1), \{S : |S| \geq m_{\mathcal{H}}(\epsilon, \delta), S \sim \mathcal{D}^m\}$
$\mathbb{P}\{\exists h_S = A(S) \in \mathcal{H}, L_{\mathcal{D}}(h_S) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon\} \geq 1 - \delta$

Chapter4 $\sim$ Chapter5

Definition 4.1 (ϵ-representative sample). A training set S is called ϵ-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$) if

$$\forall h \in \mathcal{H}, \ \ |L_S(h) - L_{\mathcal{D}}(h)| \le \epsilon.$$

Definition 4.3 (Uniform Convergence). We say that a hypothesis class $\mathcal{H}$ has the *uniform convergence property* (w.r.t. a domain Z and a loss function ℓ) if there exists a function $m_{\mathcal{H}}^{\mathrm{UC}} : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$ and for every probability distribution $\mathcal{D}$ over Z, if S is a sample of $m \ge m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon, \delta)$ examples drawn i.i.d. according to $\mathcal{D}$, then, with probability of at least $1-\delta$, S is ϵ-representative.

A hypothesis class $\mathcal{H}$ is *Uniform Convergence* if:
$\exists m_{\mathcal{H}}^{UC}(\delta, \epsilon) \to \mathbb{N}$ satisfies:
$\forall \epsilon, \delta \in (0,1), \{S : |S| \ge m_{\mathcal{H}}^{UC}(\epsilon, \delta), S \sim \mathcal{D}^m\}$
$\mathbb{P}\{\forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \le \epsilon\} \ge 1 - \delta$

Chapter6

Definition 6.5 (VC-dimension). The VC-dimension of a hypothesis class $\mathcal{H}$, denoted VCdim($\mathcal{H}$), is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size we say that $\mathcal{H}$ has infinite VC-dimension.

Theorem 6.7 (The Fundamental Theorem of Statistical Learning). *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$ and let the loss function be the* $0-1$ *loss. Then, the following are equivalent:*

1. *$\mathcal{H}$ has the uniform convergence property.*
2. *Any ERM rule is a successful agnostic PAC learner for $\mathcal{H}$.*
3. *$\mathcal{H}$ is agnostic PAC learnable.*
4. *$\mathcal{H}$ is PAC learnable.*
5. *Any ERM rule is a successful PAC learner for $\mathcal{H}$.*
6. *$\mathcal{H}$ has a finite VC-dimension.*