Machine Learning and Data Analysis
Infinite Hypothesis Spaces (wrap-up)

Filip Železný

Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Cybernetics
Intelligent Data Analysis lab
http://ida.felk.cvut.cz

January 6, 2012

Chernoff bound

Let P_Z be distribution on $\{0,1\}$ and $\{z_1, z_2, \ldots z_m\}$ be an i.i.d. sample from P_Z. Then for the difference between the true and sample means it holds

$$\Pr\left(\sum_{i=1}^{m} z_i < (1-\gamma)P_Z(1)m\right) \leq e^{-m\gamma^2/2}$$

Similar to the Hoeffding inequality we know but multiplicative.

Consequence: given an i.i.d x-sample S with m elements from P_X and a region $r \subseteq X$ such that $\sum_{x \in r} P_X(x) = \epsilon$

$$\Pr\left(|S \cap r| \leq \frac{1}{2}\epsilon m\right) \leq e^{-m/8}$$

PAC Learning with Infinite $\mathcal{F}$

Any hypothesis consistent with an ϵ-net is good ($e(f) < \epsilon$).

So we want to bound by δ the probability that a sample is not an ϵ-net.

Assume we bound by P the probability that a random error region r contains no sample point.

If there were a finite number of error regions (finite $|\Delta_\epsilon(c)|$), we could bound the probability that some of them contains no sample point by $|\Delta_\epsilon(c)|P$.

But $|\Delta_\epsilon(c)|$ is generally infinite.

Instead we adopt the *double-sampling* trick.

PAC Learning with Infinite $\mathcal{F}$ (cont'd)

Let S_1 and S_2 be two x-samples, each of size m. Distinguish two random events:

- A : S_1 is not an ϵ-net
- B : A happens and $|S_2 \cap r| > \frac{1}{2}\epsilon m$ for some $r \in \Delta_\epsilon(c)$

If A happens, then there is some $r \in \Delta_\epsilon(c)$ such that $\sum_{x \in r} P_X(x) \geq \epsilon$, so by the Chernoff bound

$$\Pr\left(|S \cap r| > \frac{1}{2}\epsilon m\right) = 1 - \Pr\left(|S \cap r| \leq \frac{1}{2}\epsilon m\right) > 1 - e^{-m/8}$$

assume that $m \geq \frac{8}{\epsilon}\ln 2$, then $1 - e^{-m/8} \geq \frac{1}{2}$ (keep the assumption)

So under these assumptions $\Pr(B|A) \geq \frac{1}{2}$.

PAC Learning with Infinite $\mathcal{F}$ (cont'd)

Remind:

- A : S_1 is not an ϵ-net
- B : A happens and $|S_2 \cap r| > \frac{1}{2}\epsilon m$ for some $r \in \Delta_\epsilon(c)$

$$\Pr(B) = \Pr(A \wedge B) = \Pr(B|A)\Pr(A)$$

we proved that $\Pr(B|A) \geq \frac{1}{2}$ so

$$\Pr(A) \leq 2\Pr(B)$$

So to bound $\Pr(A) \leq \delta$, we bound $\Pr(B) \leq \frac{\delta}{2}$

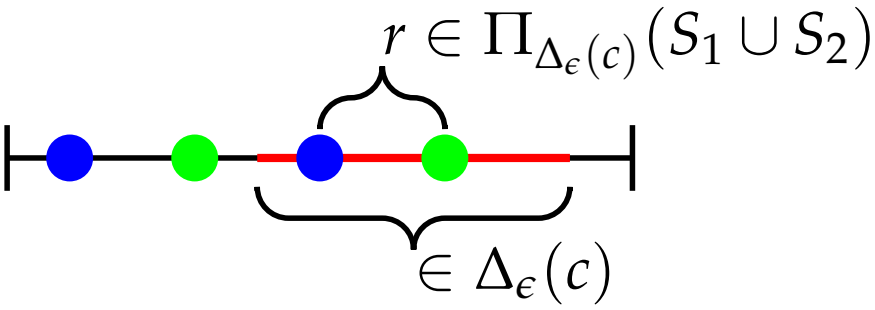
PAC Learning with Infinite $\mathcal{F}$ (cont'd)

Event B equivalently:

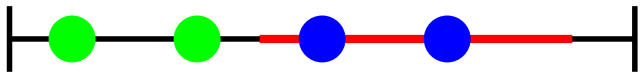
- Sample $|S_1 \cup S_2| = 2m$ points and randomly partition then into S_1 and S_2 of equal size.
- There is some $r \in \Pi_{\Delta_\epsilon(c)}(S_1 \cup S_2)$ such that

$$|r| > \frac{1}{2}\epsilon m \text{ and } r \cap S_2 = \{\}$$

Example for $m = 2$, $S_1 =$ blue, $S_2 =$ green



Event B did not occur



Event B occurred

PAC Learning with Infinite $\mathcal{F}$ (cont'd)

For a fixed r, the probability that by partitioning $S_1 \cup S_2$, all elements fall in S_2 is

$$\frac{\binom{m}{|r|}}{\binom{2m}{|r|}} \leq 2^{-|r|} \leq 2^{-\epsilon m/2}$$

(Remind that $|r| > \epsilon m/2$.)

There are $|\Pi_{\Delta_\epsilon(c)}(S_1 \cup S_2)|$ possible r's, so

$$\Pr(B) \leq |\Pi_{\Delta_\epsilon(c)}(S_1 \cup S_2)| 2^{-\epsilon m/2}$$

PAC Learning with Infinite $\mathcal{F}$ (cont'd)

$$
\begin{aligned}
\Pr(B) &\leq |\Pi_{\Delta_\epsilon(c)}(S_1 \cup S_2)| 2^{-\epsilon m/2} \\
&\leq |\Pi_{\Delta_0(c)}(S_1 \cup S_2)| 2^{-\epsilon m/2} && \text{because } \Delta_\epsilon(c) \subseteq \Delta_0(c) \\
&\leq |\Pi_{\mathcal{F}}(S_1 \cup S_2)| 2^{-\epsilon m/2} && \text{because } \mathcal{V}(\Delta_0(c)) = \mathcal{V}(\mathcal{F}) \\
&\leq G_{\mathcal{F}}(S_1 \cup S_2) 2^{-\epsilon m/2} && \text{by definition of } G_{\mathcal{F}} \\
&\leq \Phi(\mathcal{V}(\mathcal{F}), 2m) 2^{-\epsilon m/2} && \text{by the bound we proved} \\
&\leq \left(\frac{2em}{\mathcal{V}(\mathcal{F})}\right)^{\mathcal{V}(\mathcal{F})} 2^{-\epsilon m/2} && \text{plugging in the bound for } \Phi
\end{aligned}
$$

PAC Learning with Infinite $\mathcal{F}$ (cont'd)

Remind that we require

$$\Pr(B) \leq \left(\frac{2em}{\mathcal{V}(\mathcal{F})}\right)^{\mathcal{V}(\mathcal{F})} 2^{-\epsilon m/2} \leq \frac{\delta}{2}$$

Taking logarithms:

$$\mathcal{V}(\mathcal{F}) \log_2 \frac{2em}{\mathcal{V}(\mathcal{F})} - \frac{\epsilon m}{2} \leq \log_2 \frac{\delta}{2}$$

So for m:

$$m \geq \frac{2}{\epsilon} \log_2 \frac{2}{\delta} + \frac{2\mathcal{V}(\mathcal{F})}{\epsilon} \log_2 \frac{2em}{\mathcal{V}(\mathcal{F})}$$

PAC Learning with Infinite $\mathcal{F}$ (cont'd)

$$m \geq \frac{2}{\epsilon} \log_2 \frac{2}{\delta} + \frac{2\mathcal{V}(\mathcal{F})}{\epsilon} \log_2 \frac{2em}{\mathcal{V}(\mathcal{F})}$$

To guarantee this, we make $m/2$ greater than each of the two summands.

For the first summand:

$$m \geq \frac{4}{\epsilon} \log_2 \frac{2}{\delta}$$

but earlier we also assumed $m \geq \frac{8}{\epsilon} \ln 2$, so we strenghten the above:

$$m \geq \frac{8}{\epsilon} \log_2 \frac{2}{\delta} \geq \frac{8}{\epsilon} \ln 2$$

PAC Learning with Infinite $\mathcal{F}$ (cont'd)

For the second summand:

$$m \geq \frac{4\mathcal{V}(\mathcal{F})}{\epsilon} \log_2 \frac{2em}{\mathcal{V}(\mathcal{F})}$$

Left-hand side grows faster (linearly) in m than right-hand side (logarithmically). Thus we only need to find a minimum value m_0 of m satisfying the inequality, and it will be satisfied for all $m \geq m_0$.

Plugging

$$m_0 = \frac{8\mathcal{V}(\mathcal{F})}{\epsilon} \log_2 \frac{13}{\epsilon}$$

for m in the inequality, we obtain (verify!) the equivalent inequality

$$\frac{13^2}{16e\epsilon} \geq \log_2 \frac{13}{\epsilon}$$

which is true for any $\epsilon \leq 1$.

PAC Learning with Infinite $\mathcal{F}$: Result

PAC Learning with Infinite $\mathcal{F}$

Let $\mathcal{F}$ be a hypothesis class with a finite $\mathcal{V}(\mathcal{F})$ and $\mathcal{C}$ be concept class, both on X. Let $c \in \mathcal{C}$ be a concept. A hypothesis f consistent with a sample $\{(x_1, c(x_1)), \ldots, (x_m, c(x_m))\}$ will have $e(f) \leq \epsilon$ with probability at least $1 - \delta$ if

$$m \geq \max\left(\frac{8}{\epsilon}\log_2\frac{2}{\delta}\,,\ \frac{8\mathcal{V}(\mathcal{F})}{\epsilon}\log_2\frac{13}{\epsilon}\right)$$

Therefore any $\mathcal{C}$ is (efficiently) PAC-learnable by $\mathcal{F}$ if there is an (efficient) learner producing a consistent $f \in \mathcal{F}$ for any sample, and $\mathcal{V}(\mathcal{F})$ is polynomial (in the size of examples n).

As we have seen, $\mathcal{V}(\mathcal{F})$ is usually linear in the number of hypothesis class parameters, which corresponds to n.

$\mathcal{V}(\mathcal{F})$: Remarks

- The result can be rewritten into a simpler form

$$m \geq c_0 \left(\frac{\mathcal{V}(\mathcal{F})}{\epsilon} \log_2 \frac{1}{\epsilon} + \frac{1}{\epsilon} \log_2 \frac{1}{\delta} \right)$$

  where c_0 is a constant.
- The result holds also for finite $\mathcal{F}$. For some $\mathcal{F}$, it may even provide better bounds than those we derived specially for finite $\mathcal{F}$.
- Finite $\mathcal{V}(\mathcal{F})$ is also a necessary condition for PAC-learning. It can be proved that at least

$$\frac{\mathcal{V}(\mathcal{F}) - 1}{64\epsilon}$$

  examples are needed to PAC-learn a concept class with $\mathcal{F}$ if $\delta \leq 1/15$.

Error Bounds for Infinite $\mathcal{F}$

$\mathcal{V}(\mathcal{F})$ also enables to derive error bounds for inconsistent hypotheses.
$\mathcal{V}(\mathcal{F})$ is 'analogical' to $\ln|\mathcal{F}|$ for finite hypothesis classes.

With probability at least $1-\delta$, for a training set S:

$$|e(f) - e(S,f)| \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{V}(\mathcal{F})}{m}\log_2\frac{m}{\mathcal{V}(\mathcal{F})} + \frac{1}{m}\log_2\frac{1}{\delta}}\right)$$

and if f minimizes training error $e(f,S)$ then with probability at least $1-\delta$:

$$e(f) \leq e(f^*) + \mathcal{O}\left(\sqrt{\frac{\mathcal{V}(\mathcal{F})}{m}\log_2\frac{m}{\mathcal{V}(\mathcal{F})} + \frac{1}{m}\log_2\frac{1}{\delta}}\right)$$

where f^* minimizes classification error $e(f)$.

Universal learnability

PAC bounds may be very loose (pessimistic) in practice due to overly general assumptions.

Prior probabilities of hypotheses are not assumed in the PAC model.

The *universal learnability* framework (Muggleton, Page 1997) is an extension of the PAC-model where a prior probability distribution $P_{\mathcal{F}}(f)$ on hypotheses is considered. For example:

$$P_{\mathcal{F}}(f) \approx e^{-\mathsf{DL}(f)}$$

where $\mathsf{DL}(f)$ is the description length (i.e., complexity) of the hypothesis in a suitable language. Many real-life learning algorithms indeed prefer simpler hypotheses.

The Universal learnability framework enables to prove some stronger results than the PAC-framework.

Bayesian PAC-learning

Bayesian PAC-learning is a currently vivid stream in computational learning theory.

Prior probabilities of hypotheses are also assumed as in Universal learnability. After seeing training sample S, they are used to derive the *posterior probabilities*

$$P_{\mathcal{F}|\mathcal{S}}(f|S) \approx P_{\mathcal{S}|\mathcal{F}}(S|f)P_{\mathcal{F}}(f) \approx e(f,S)P_{\mathcal{F}}(f)$$

A new example x is then classified into class 1 iff

$$\sum_{f \in \mathcal{F}} f(x) P_{\mathcal{F}|\mathcal{S}}(f|S) > \frac{1}{2}$$

(This decision be approximated by sampling a set of f from $P_{\mathcal{F}|\mathcal{S}}$ and taking the majority vote.)

Stronger results may be derived than in the conventional PAC-framework.

Different Learning Protocols

The conventional PAC-learning framework applies to the supervised learning protocol. It does not cover other protocols such as

Learning with queries. The learning algorithm may ask queries to the teacher (oracle). Existing frameworks consider e.g.:

- membership queries, such as *what is the class of* x?
- statistical queries, such as *what is the probability of drawing* x *such that* $x^{(1)} \in [0.3, 0.4]$*?*

The statistical query model proves some stronger results (i.e., fewer examples needed) than the PAC model.

Reinforcement learning. Instead of class information, feedback from the teacher is received after a sequence of decisions made by the learner.

Other scenarios. Such as those where decisions made by the learner influence $P_{Y|X}$.

Bias-Variance Trade-off Revisited

Remind: in the finite $\mathcal{F}$ case, by extending $\mathcal{F}$

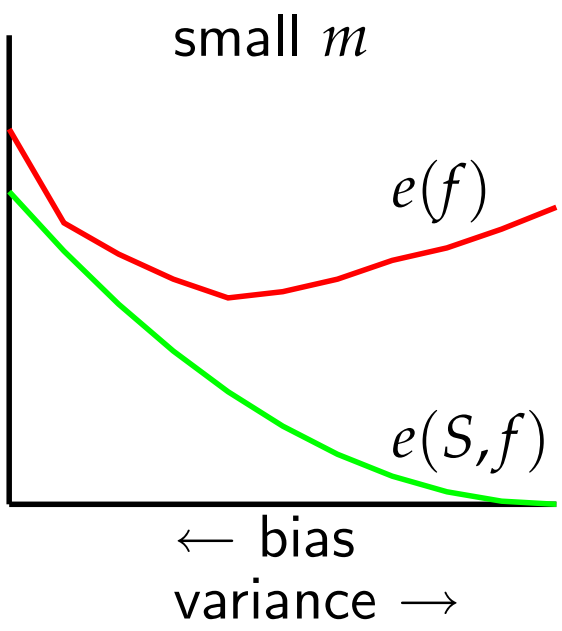
$$e(f) \leq \underbrace{\left(\min_{f \in \mathcal{F}} e(f)\right)}_{\text{'bias': may decrease}} + \underbrace{2\sqrt{\frac{1}{2m} \ln \frac{2|\mathcal{F}|}{\delta}}}_{\text{'variance': will increase}}$$

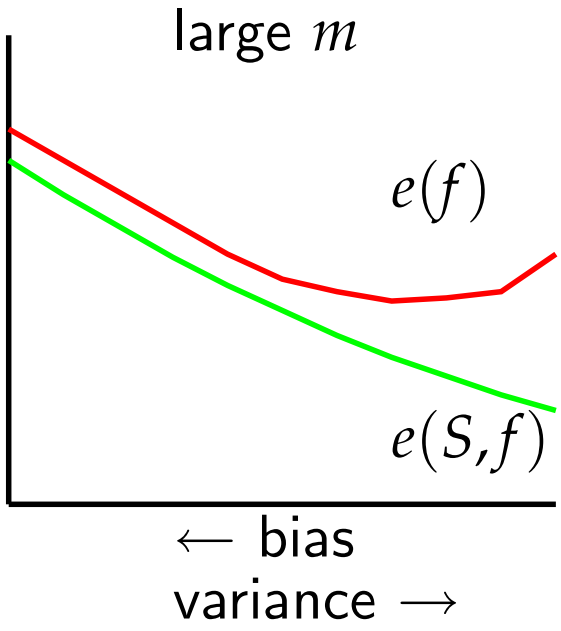
This holds analogically for infinite $\mathcal{F}$

$$e(f) \leq \underbrace{\left(\min_{f \in \mathcal{F}} e(f)\right)}_{\text{'bias': may decrease}} + \underbrace{\mathcal{O}\left(\sqrt{\frac{\mathcal{V}(\mathcal{F})}{m} \log_2 \frac{m}{\mathcal{V}(\mathcal{F})} + \frac{1}{m} \log_2 \frac{1}{\delta}}\right)}_{\text{'variance': will increase}}$$

Bias-Variance Trade-off Revisited (cont'd)

Resulting behavior (we have seen this before)





Bias-Variance Trade-off in Regression

The PAC framework provides bounds for classification. Not applicable in regression.

An analogy of the bias-variance trade-off may be derived for regression. Consider loss function

$$L(y, y') = L_{SQ}(y, y') = (y - y')^2$$

Under L_{SQ}, risk of f is the mean squared error

$$MSE(f) = \int_{x \in X} \int_{y \in Y} L_{SQ}\left(y, f(x)\right) \mathrm{d}P_{XY}(x, y)$$

Bias-Variance Trade-off in Regression (cont'd)

$MSE(f)$ can be expressed as

$$MSE(f) = \int_{x \in X} MSE(f(x)) \mathrm{d}P_X(x)$$

where

$$MSE(f(x)) = \int_{y \in Y} L_{SQ}\left(y, f(x)\right) \mathrm{d}P_{Y|X}(y|x)$$

is the mean squared error at a *fixed* x.

For simplicity, we assume that $P_{Y|X}$ is deterministic, i.e.

$$P_{Y|X}(y|x) = 1 \text{ iff } f^*(x) = y$$

where $f^* : X \to Y$ is some function. Then

$$MSE(f(x)) = (f(x) - f^*(x))^2$$

Bias-Variance Trade-off in Regression (cont'd)

In $MSE(f(x))$, f is learned from a given sample S. We now study how $f(x)$ behaves over multiple samples S.

S is random (drawn from $\mathcal{S} = 2^X$), therefore $f(x)$ is now random even though x is fixed. Then also $MSE(f(x))$ is a random variable and for its mean value we have:

$$
\begin{aligned}
&\mathbf{E}_{\mathcal{S}}\left[MSE(f(x))\right] = \mathbf{E}_{\mathcal{S}}\left[\left(f(x) - f^*(x)\right)^2\right] \\
=&\mathbf{E}_{\mathcal{S}}\left[f(x)^2 - 2f(x)f^*(x) + f^*(x)^2\right] \\
=&\mathbf{E}_{\mathcal{S}}\left[f(x)^2\right] - 2\mathbf{E}_{\mathcal{S}}\left[f(x)\right]f^*(x) + f^*(x)^2 + \mathbf{E}_{\mathcal{S}}\left[f(x)\right]^2 - \mathbf{E}_{\mathcal{S}}\left[f(x)\right]^2 \\
=&\mathbf{E}_{\mathcal{S}}\left[f(x)^2\right] - \mathbf{E}_{\mathcal{S}}\left[f(x)\right]^2 + \left(\mathbf{E}_{\mathcal{S}}\left[f(x)\right] - f^*(x)\right)^2 \\
=&\mathrm{Var}_{\mathcal{S}}\left(f(x)\right) + \mathrm{Bias}^2_{\mathcal{S}}(f^*(x), f(x))
\end{aligned}
$$

(note the linearity of the expectation operator $\mathbf{E}$)

Bias-Variance Trade-off in Regression (cont'd)

More generally, assume that $P_{Y|X}$ is not deterministic. In particular, assume

$$y = f^*(x) + z$$

where z is a random variable ('additive noise'). Then

$$\mathbf{E}\left[MSE(f(x))\right] = \mathrm{Var}_{\mathcal{S}}\left(f(x)\right) + \mathrm{Bias}^2_{\mathcal{S}}(f^*(x), f(x)) + \mathrm{Var}_{\mathcal{S}}\left(z\right)$$

Here, $\mathrm{Var}_{\mathcal{S}}\left(z\right)$ is the *irreducible error*.