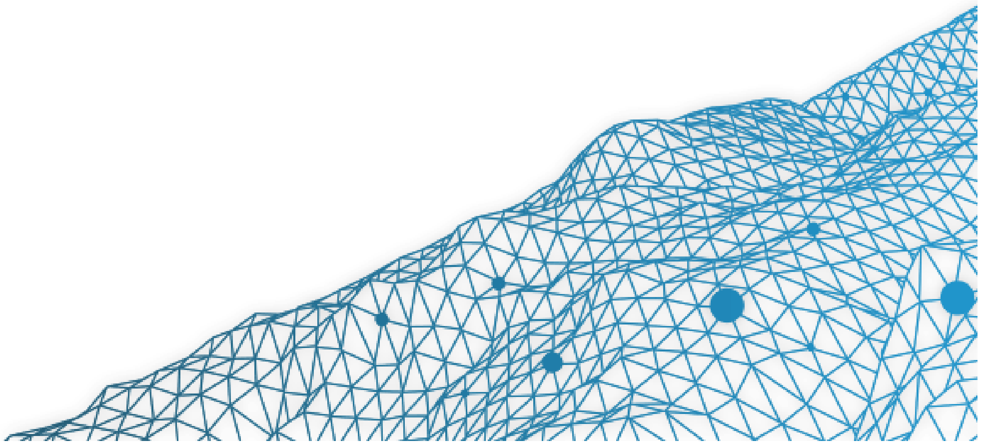
7 Classification & Goodness of Fit (Theoretical)

Arthur Charpentier (Université du Québec à Montréal)

Machine Learning & Econometrics

SIDE Summer School - July 2019



Machine Learning, in a Probabilistic Framework

Assume that training and validation data are drawn i.i.d. from $\mathbb{P}$, or $(Y, \boldsymbol{X}) \sim F$

Consider $y \in \{-1, +1\}$. The true risk of a classifier is

$$\mathcal{R}(m) = \mathbb{P}_{(Y,\boldsymbol{X})\sim F}\big(m(\boldsymbol{X}) \neq Y)\big) = \mathbb{E}_{(Y,\boldsymbol{X})\sim F}\big(\ell(m(\boldsymbol{X}), Y)\big)$$

Bayes classifier is

$$b(\boldsymbol{x}) = \text{sign}\big(\mathbb{E}_{(Y,\boldsymbol{X})\sim F}\big[Y|\boldsymbol{X} = \boldsymbol{x}\big]\big)$$

which satisfies $\mathcal{R}(b) = \inf_{m\in\mathcal{H}} \{\mathcal{R}(m)\}$ (in the class $\mathcal{H}$ of all measurable functions), called Bayes risk.

The empirical risk is

$$\widehat{\mathcal{R}}_n(m) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, m(\boldsymbol{x}_i))$$

Machine Learning, in a Probabilistic Framework

One might think of using regularized empirical risk minimization,

$$\widehat{m}_n \in \underset{m\in\mathcal{M}}{\text{argmin}}\left\{\widehat{\mathcal{R}}_n(m) + \lambda\|m\|\right\}$$

in a class of models $\mathcal{M}$, where regularization term will control the complexity of the model to prevent overfitting.

Let $m^\star$ denote the best model in $\mathcal{M}$, $m^\star = \underset{m\in\mathcal{M}}{\text{argmin}}\{\mathcal{R}(m)\}$

$$\mathcal{R}(\widehat{m}_n) - \mathcal{R}(b) = \underbrace{\mathcal{R}(\widehat{m}_n) - \mathcal{R}(m^\star)}_{\text{estimation error}} + \underbrace{\mathcal{R}(m^\star) - \mathcal{R}(b)}_{\text{approximation error}}$$

Since $\mathcal{R}(\widehat{m}_n) = \widehat{\mathcal{R}}_n(\widehat{m}_n) + [\mathcal{R}(\widehat{m}_n) - \widehat{\mathcal{R}}_n(\widehat{m}_n)]$, we can write

$$\mathcal{R}(\widehat{m}_n) \leq \widehat{\mathcal{R}}_n(\widehat{m}_n)) + \text{something}(m, \mathcal{M})$$

Machine Learning, in a Probabilistic Framework

To quantify this something$(m,\mathcal{F})$, we need Hoeffding inequality, see Hoeffding (1963, Probability inequalities for sums of bounded random variables)

Let $g(\boldsymbol{x},y)=\ell(m(\boldsymbol{x}),y)$, for some model m. Let

$$\mathcal{G}=\{g:(\boldsymbol{x},y)\mapsto\ell(m(\boldsymbol{x}),y),m\in\mathcal{M}\}$$

If $Z=(Y,\boldsymbol{X})$, set $\mathcal{R}(g)=\mathbb{E}_{Z\sim F}\big(g(Z)\big)$ and $\widehat{\mathcal{R}}_n(g)=\dfrac{1}{n}\sum_{i=1}^n g(z_i)$.

Hoeffding inequality

If $Z_1,\cdots,Z_n$ are i.i.d. and if h is a bounded function (in $[a,b]$), then, $\forall\epsilon>0$

$$\mathbb{P}_n\left[\left|\frac{1}{n}\sum_{i=1}^n h(Z_i)-\mathbb{E}_F[h(Z)]\right|\geq\epsilon\right]\leq 2\exp\left(\frac{-2n\epsilon^2}{(b-a)^2}\right)$$

Machine Learning, in a Probabilistic Framework

or equivalently (let δ denote the upper bound)

$$\mathbb{P}_n\left[\left|\frac{1}{n}\sum_{i=1}^n h(Z_i)-\mathbb{E}_F[h(Z)]\right|\geq (b-a)\sqrt{\frac{-1}{2n}\log(2\delta)}\right]\leq\delta$$

We can actually derive a one side majoration, and with probability (at least) $1-\delta$

$$\mathcal{R}(g)\leq\widehat{\mathcal{R}}_n(g)+\sqrt{\frac{-1}{2n}\log\delta}$$

$\mathcal{R}(m)-\widehat{\mathcal{R}}_n(m)$

For a fixed $m\in\mathcal{M}$, $\mathcal{R}(m)-\widehat{\mathcal{R}}_n(m)\sim\dfrac{1}{\sqrt{n}}$

But it doesn't help much, we need uniform deviations (or worst deviation).

Machine Learning, in a Probabilistic Framework

Consider a finite set of models. Define the set of bad samples

$$\mathcal{Z}_j = \left\{(z_1, \cdots, z_n) : \mathcal{R}(g_j) - \widehat{\mathcal{R}}_n(g_j) \geq 0\right\}$$

so that $\mathbb{P}[(Z_1, \cdots, Z_n) \in \mathcal{Z}_j] \leq \delta$, and then

$$\mathbb{P}[(Z_1, \cdots, Z_n) \in \mathcal{Z}_1 \cap \mathcal{Z}_1] \leq \mathbb{P}[(Z_1, \cdots, Z_n) \in \mathcal{Z}_1] + \mathbb{P}[(Z_1, \cdots, Z_n) \in \mathcal{Z}_2] \leq 2\delta$$

so that

$$\mathbb{P}\left[(Z_1, \cdots, Z_n) \in \bigcap_{j=1}^{\nu} \mathcal{Z}_j\right] \leq \sum_{j=1}^{\nu} \mathbb{P}[(Z_1, \cdots, Z_n) \in \mathcal{Z}_j] \leq \nu\delta$$

Hence,

$$\mathbb{P}[\exists g \in \{g_1, \cdots, g_\nu\} : \mathcal{R}(g) - \widehat{\mathcal{R}}_n(g) \geq \epsilon] \leq \nu \cdot \mathbb{P}[R(g) - \widehat{R}_n(g) \geq \epsilon] \leq \nu \cdot \exp[-2n\epsilon^2]$$

Machine Learning, in a Probabilistic Framework

If $\delta = \nu \exp[-2n\epsilon^2]$, we can write ϵ and with probability (at least) $1-\delta$

$$\forall g \in \{g_1, \cdots, g_\nu\}, \mathcal{R}(g) - \widehat{\mathcal{R}}_n(g) \leq \sqrt{\frac{1}{n}(\log \nu - \log \delta)}$$

Thus, we can write, for a finite set of models $\mathcal{M} = \{m_1, \cdots, m_\nu\}$,

$$\forall m \in \{m_1, \cdots, m_\nu\}, \mathcal{R}(m) \leq \widehat{\mathcal{R}}_n(m) + \sqrt{\frac{1}{n}(\log \nu - \log \delta)}$$

$\mathcal{R}(m) - \widehat{\mathcal{R}}_n(m)$ - $\mathcal{M}$ finite, $\nu = |\mathcal{M}|$

For the worst case scenario

$$\sup_{m \in \mathcal{M}_\nu} \left\{\mathcal{R}(m) - \widehat{\mathcal{R}}_n(m)\right\} \sim \frac{\log \nu}{\sqrt{n}}$$

Now, what if $\mathcal{M}$ is infinite ?

Machine Learning, in a Probabilistic Framework

Write Hoeffding's inequality as

$$\mathbb{P}\left[\mathcal{R}(g)-\widehat{\mathcal{R}}_n(g)\geq\sqrt{\frac{-1}{2n}\log\delta_g}\right]\leq\delta_g$$

so that, we a countable set $\mathcal{G}$

$$\mathbb{P}\left[\exists g\in\mathcal{G}:\mathcal{R}(g)-\widehat{\mathcal{R}}_n(g)\geq\sqrt{\frac{-1}{2n}\log\delta_g}\right]\leq\sum_{g\in\mathcal{G}}\delta_g$$

If $\delta_g=\delta\cdot\mu(g)$ where μ is some measure on $\mathcal{G}$, with probability (at least) $1-\delta$,

$$\forall g\in\mathcal{G},\mathcal{R}(g)\leq\widehat{\mathcal{R}}_n(g)+\sqrt{\frac{-1}{2n}[\log\delta+\log\mu(g)]}$$

(see previous computations with $\mu(g)=\nu^{-1}$)

Machine Learning, in a Probabilistic Framework

More generally, given a sample $\boldsymbol{z} = \{z_1, \cdots, z_n\}$, let $\mathcal{M}_{\boldsymbol{z}}$ denote the set of classification that can be obtained,

$$\mathcal{M}_{\boldsymbol{z}} = \big\{(m(z_1), \cdots, m(z_n))\big\}$$

The growth function is the maximum number of ways into which n points can be classified by the function class $\mathcal{M}$

$$G_{\mathcal{M}}(n) = \sup_{\boldsymbol{z}} \{\mathcal{M}_{\boldsymbol{z}}\}$$

Vapnik-Chervonenkis : with (at least) probability $1-\delta$,

$$\forall m \in \mathcal{M}, \mathcal{R}(m) \leq \widehat{\mathcal{R}}_n(m) + 2\sqrt{\frac{2}{n}[\log G_{\mathcal{M}}(2n) - \log(4\delta)]}$$

The VC (Vapnik-Chervonenkis) dimension is the largest n such that $G_{\mathcal{M}}(n) = 2^n$. It will be denoted $\text{VC}(\mathcal{M})$. Observe that $G_{\mathcal{M}}(n) \leq 2^n$

Machine Learning, in a Probabilistic Framework

$n \leq \text{VC}(\mathcal{M})$: $n \mapsto G_{\mathcal{M}}(n)$ increases exponentially $G_{\mathcal{M}}(n) = 2^n$

$n \geq \text{VC}(\mathcal{M})$: $n \mapsto G_{\mathcal{M}}(n)$ increases at power speed $G_{\mathcal{M}}(n) \leq \left(\dfrac{en}{\text{VC}(\mathcal{M})}\right)^{\text{VC}(\mathcal{M})}$

Vapnik-Chervonenkis : with (at least) probability $1-\delta$,

$$\forall m \in \mathcal{M}, \mathcal{R}(m) \leq \widehat{\mathcal{R}}_n(m) + 2\sqrt{\frac{2}{n}\left[\text{VC}(\mathcal{M})\log\left(\frac{en}{\text{VC}(\mathcal{M})}\right) - \log(4\delta)\right]}$$

$\mathcal{R}(m) - \widehat{\mathcal{R}}_n(m)$ - $\mathcal{M}$ infinite

For the worst case scenario $\displaystyle\sup_{m\in\mathcal{M}}\left\{\mathcal{R}(m) - \widehat{\mathcal{R}}_n(m)\right\} \sim \sqrt{\frac{\text{VC}(\mathcal{M})\cdot\log n}{n}}$

To go further, see Bousquet, Boucheron & Lugosi (2005, Introduction to Learning Theory)