Lecture Notes V – Model selection

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

November, 2020

Cross-validation

AIC and BIC

Structural risk minimization and VC dimension

Reading HTF Ch.: Ch. 7, Murphy Ch.: BIC, AIC 8.4.2 (pp 255), SRM 6.5 (pp 204)

Crossvalidation

K-fold Crossvalidation

AIC and BIC

Hold for

- parametric $\mathcal{F}$
- log-likelihood loss $L(y, f(x)) = -\ln P(y|x, f)$
  Note that $-N\hat{L}(f) = \ln P(y^{1:n}|x^{1:n}, f)$ data log likelihood
- $\hat{f} \in \mathcal{F}$ estimated by Maximum Likelihood
- (for BIC: $\frac{\partial^2 L}{\partial \text{parameters}}$ non-singular at $\hat{f}$)

Akaike's Information Criterion (AIC)

$$AIC(\hat{f}) = -N\hat{L}(\hat{f}) - d, \tag{1}$$

where $d = \#\text{parameters}(f)$, and $N =$ the size of $\mathcal{D}$.

The Bayesian Information Criterion (BIC)

$$BIC(\hat{f}) = -N\hat{L}(\hat{f}) - \frac{d}{2}\ln N, \tag{2}$$

with $d = \#\text{parameters}(f)$

AIC and BIC

VC dimension

$\mathcal{F}$ shatters $\mathcal{D}_h = \{x^1, \ldots x^h\}$

iff, for every possible labeling $y^{1:h} \in \{\pm 1\}$ of $\mathcal{D}_h$, there is a function $f \in \mathcal{F}$ that achieves that labeling, i.e. $\mathrm{sgn} f(x^i) = y^i$ for all $i = 1 : m$.

VC dimension

$\mathcal{F}$ shatters $\mathcal{D}_h = \{x^1, \ldots x^h\}$

iff, for every possible labeling $y^{1:h} \in \{\pm 1\}$ of $\mathcal{D}_h$, there is a function $f \in \mathcal{F}$ that achieves that labeling, i.e. $\mathrm{sgn} f(x^i) = y^i$ for all $i = 1 : m$.

VC dimension of $\mathcal{F}$

A model class $\mathcal{F}$ over $\mathbb{R}^n$ has VC dimension h iff h is the maximum positive integer so that there exists a set of h points in $\mathbb{R}^n$ that is shattered by $\mathcal{F}$.

Structural risk minimization

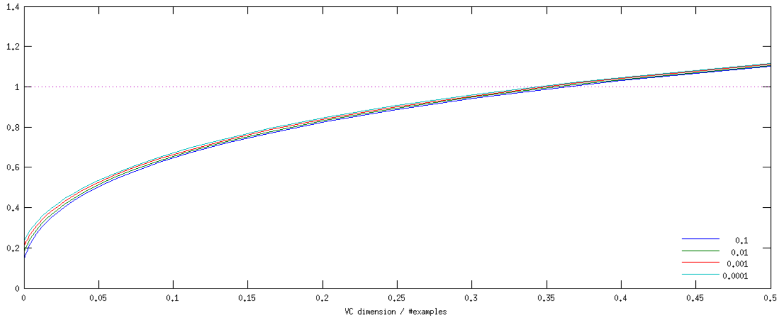
Theorem

Let $\mathcal{F}$ be a model class of VC-dimension h and f a classifier in $\mathcal{F}$. Then, with probability w.p. $> 1 - \delta$ over training sets

$$L_{01}(f) \;\leq\; \hat{L}_{01}(f) + \sqrt{\frac{h[1 + \log(2N/h)] + \log(4/\delta)}{N}}. \tag{3}$$

Structural risk minimization

$$\sqrt{\frac{h[1+\log(2N/h)]+\log(4/\delta)}{N}}$$



Structural risk minimization

Theorem

Let $\mathcal{F}$ be a model class of VC-dimension h, with $f(x) \in [-1, 1]$ for all x and for all $f \in \mathcal{F}$. Let $\delta > 0$ and $\zeta \in (0, 1)$. Denote $\mathcal{D} = \{(x^i, y^i), i = 1 : N\}$ the current training set. Then, with probability w.p. $> 1 - \delta$ over training sets

$$L_{01}(f) \;\leq\; \hat{L}_{01,\zeta}(f) + \tilde{\mathcal{O}}\left(\sqrt{\frac{h}{N\zeta^2}}\right) \tag{4}$$

for any $f \in \mathcal{F}$.

The test set method of bounding the classification error

Given a classifier f and a data set $\mathcal{D}^{\text{test}}$ of size N.

$\hat{L}_{01}(f) \sim Binomial(L_{01}(f), N)$ denote $\bar{b}(m, L_{01}, \delta) = \max\{L_{01} \mid Pr[m \mid L_{01}, N] \geq \delta\}$

$$L_{01}(f) \leq \hat{L}_{01}(f) + \sqrt{\frac{\ln 1/\delta}{2N}} \quad \text{w.p. } 1-\delta \tag{5}$$

$$L_{01}(f) \leq \hat{L}_{01}(f) + \sqrt{\frac{2\hat{L}_{01}\ln 1/\delta}{N} + \frac{2\ln(1/\delta)}{N}} \quad \text{w.p. } 1-\delta \tag{6}$$

$$L_{01}(f) \leq \frac{\ln 1/\delta}{N} \quad \text{w.p. } 1-\delta \text{ when } \hat{L}_{01}(f) = 0 \tag{7}$$

$$|L_{01}(f) - \hat{L}_{01}(f)| \leq \sqrt{\frac{\ln 1/\delta}{2N}} \quad \text{w.p. } 1-\delta \tag{8}$$

The test set method of bounding the classification error

$|L_{01}(f) - \hat{L}_{01}(f)| \leq \sqrt{\frac{\ln 1/\delta}{2N}}$ w.p. $1 - \delta$