STAT/BIOST 527 Lecture III Notes

April, 2023

# Brief overview of Model Selection and Regularization

©Marina Meilă

mmp@stat.washington.edu

ONLY CV relevant to STAT/BIOST 527

*Reading:* Murphy: BIC, AIC 8.4.2 (pp 255), SRM 6.5 (pp 204) Hastie, Tibshirani & Friedman: chapter 7, John Langford, *"Tutorial on practical prediction theory for classification"*, JMLR **6**, (2005), 273–306 (section 3)

## 1 Cross-Validation

The idea of cross-validation is to "test" a trained model on "fresh" data, data that has not been used to construct the model. Of course, we need to have access to such data, or to set aside some data before building the model. This data set is called *validation data* or *hold out* data (or sometimes *test data*, in contrast to the data used to build the model which is called *training data*.

We will "validate" the model on the holdout data. If the model is accurate, it must be able to predict well unseen data coming from the same distribution. The loss of our trained predictor on the test data

$$L^{CV}(f) \quad = \quad \frac{1}{|\mathcal{D}'|} \sum_{x \in \mathcal{D}'} L(y, f(x)) \tag{1}$$

is a measure of the goodness of our predictor. If we have fitten several predictors $f^{1:m}$ to a training set $\mathcal{D}$, we compare the quality of their predictions on $\mathcal{D}'$, and we choose

$$f^* \; = \; \operatorname*{argmin}_{1:m} L^{CV}(f^j) \tag{2}$$

as the "best" predictor.

In prediction problems, $L$ stands for prediction loss, for example Misclassfication Error or Least Squares Error; in density estimation it represents the log-likelihood.

**The size of the validation set $\mathcal{D}_{test}$.** If the validation set is too small, then the value of $L^{CV}$ will have high variance (i.e will change much if we pick another validation set out of the original data set). So, our decision based on it will be prone to error. But if $n' = |\mathcal{D}'|$ is large, then we may be left with too little data for building the model. To balance the two, consider that $\mathcal{D}'$ will be used for estimating only one number, its loss, for each $f^j$, while $\mathcal{D}$ will be used to estimate all the parameters of $f^j$. Hence, the rule of thumb when sufficient data is available is to choose a set of $n' = 200 \dots 1000$ samples for validation, and to use the remaining ones for training.

For smaller data sets, a procedure called **$K$-fold cross validation** is used. The whole data is divided at random into equal sized sets $\mathcal{D}_1, \dots \mathcal{D}_K$. Then, for $k = 1 \dots K$, $\mathcal{D}_k$ is used as a validation set, while the rest of the data is used as training set. The loss $L_k(f^j)$ of $\mathcal{D}_k$ for the $j$-th model (trained on the remaining data) is calculated. The final score for each $f^j$ is equal to the arithmetic mean of $L^k(f^j)$, $k = 1 \dots K$. In practice, the values of $K$ range from 5–10 to $n = |\mathcal{D}|$. If $K = n$ the method is called **leave-one-out cross validation**.

$K$-fold CV in more detail, in the case of kernel density estimation. In this case, each model class $j$ is represented by a bandwidth $h$.

- Partition the data $\mathcal{D}$ into $K$ disjoint sets of equal size $n/K$, $\mathcal{D}_{1:K}$.
- for each $h$
  1. for each $k = 1 : K$
     (a) train predictor/density estimator on $\mathcal{D}_{-k} = \mathcal{D} \setminus \mathcal{D}_k$, obtain $p^k_{X,h}$
     (b) calculate $L^k(h) = \frac{K}{n} \sum_{i \in \mathcal{D}_k} \ln p^k_{X,h}(x^i)$
  2. average $L^k(h)$ to obtain $\bar{L}^k(h)$ and its standard deviation $\sigma_L(h)$.

Rule 1 $h^{opt} = \underset{h}{\text{argmax/min}} \bar{L}(h)$, $L^{opt} = L(h^{opt})$ select $h$ that minimizes Loss/maximizes log-likelihood;
OR

Rule 2 $h^{opt2} = \max / \min\{h; \bar{L}(h) \geq L^{opt} - \sigma_L(h^{opt})\}$. Of all $h$ that have $L(h)$ within a standard deviation of the maximum, select the one that corresponds to the *smoothest* model.

K-fold CV is computationally expensive, however costs vary relatively little with $K$.

## 2 AIC and BIC

Assume that the loss function is based on a likelihood (i.e. for predictors $f$ associated with a probabilistic model), i.e. $L(y, f(x)) = -\ln P(y|x, f)$, and that $f$ is fit by Maximum Likelihood (i.e. by minimizing $\hat{L}$. Then, we have we have two criteria for model selection that use the data only through $\hat{L}$.

**Akaike's Information Criterion (AIC)** is defined as

$$AIC(f) \;=\; -n\hat{L}(f) - d, \tag{3}$$

where $d = \#\text{parameters}(f)$. We select the model $f$ for which

$$AIC(f) \;=\; \max_{j=1:m} AIC(f^j) \tag{4}$$

Thus, AIC penalizes *likelihood* on the training set with the number of parameters used to achieve this fit.

**The Bayesian Information Criterion (BIC)** is applies in similar conditions to AIC.

$$BIC(f) \;=\; -n\hat{L}(f) - \frac{d}{2}\ln n, \tag{5}$$

with $d = \#\text{parameters}(f)$ and $n =$ the sample size of the $\mathcal{D}$ used to fit $f$. As with AIC, the predictor (model) with the maximum BIC is selected as "best".

Remark: AIC and BIC are used in a more general framework than prediction, namely where a log-likelihood is maximized to obtain a parametrized model[1].

It is obvious that for all but the very smallest sample sizes $n$, BIC will penalize a model more than AIC, therefore AIC will choose models with more (or the same number of) parameters than BIC. Why this difference? The difference comes from the principles that are at the basis of AIC and BIC. The former is an (asymptotic) estimator of the *expected loss* of a model (and asymptotically will behave the same as leave-one-out CV), whereas the latter is an (asymptotic) estimator of the *marginal likelihood* of a model given the data.

[1] The model must satisfy certain *regularity conditions*. A remarkable case where BIC does not apply is the case of statistical models with latent (unobserved) variables; in this case the BIC can over-penalize the model.

Specifically, for a model family $\mathcal{F} = \{f_\theta, \theta \in \Theta \subseteq \mathbb{R}^d\}$, BIC approximates $P(\mathcal{F}|\mathcal{D}) = \int_\Theta Prior(f) \prod_{(x,y)\in\mathcal{D}} P(y|x, f_\theta) d\theta$.

Hence, AIC selects the models that predicts best, while BIC selects the model that is the best explanation of the data. On finite samples, these two are often not the same.

# 3 Structural Risk Minimization

## 3.1 VC dimension and model complexity

Model complexity, even for a parametric model, is not always the same as the number of free parameters in the model. Model complexity is a task dependent measure. For classification, an important measure of complexity is the *Vapnik-Chervonenkis (VC) dimension* of a model class $\mathcal{F}$.

**Definition 1** *We say that model class $\mathcal{F}$* **shatters** *a set of points $\mathcal{D}_h = \{x^1, \ldots x^h\}$ 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.* $\operatorname{sign} f(x^i) = y^i$ *for all $i = 1 : m$.*

**Definition 2** *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 $\mathcal{F}$.*

The VC-dimension is always an integer, but its value for a model class $\mathcal{F}$ can rarely be calculated or estimated. (E.g. we don't know the VC dimension of neural networks.)

**Example 1** *The family of linear classifiers $\mathcal{F} = \{f(x) = w^T x + b, x, w \in \mathbb{R}^n, b \in \mathbb{R}\}$ has $VCdim$ equal to $n + 1$.*

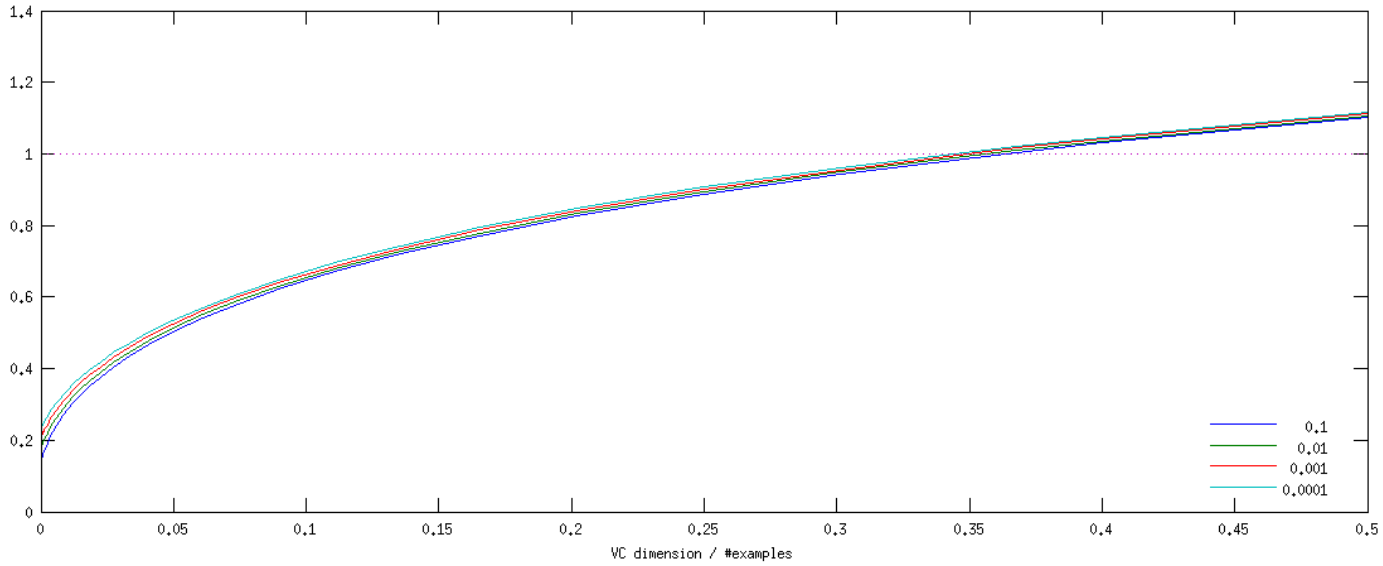
## 3.2 Structural risk minimization (SRM)

The importance of the VC-dimension comes from theorems such as this one.

**Theorem 1** *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*

$$L_{01}(f) \ \leq \ \hat{L}_{01}(f) + \sqrt{\frac{h[1 + \log(2n/h)] + \log(4/\delta)}{n}}. \tag{6}$$

In other words, for a small VC-dimension, the empirical loss $\hat{L}_{01}(f)$ is a good predictor of the true loss $L_{01}(f)$. The figure below displays graphically the last term from equation (6) for values of $\delta = 0.1, \ldots 0.0001$.




Like AIC and BIC, equation (6) balances $\hat{L}$ with a penalty that depends on the model complexity and $n$. But, unlike AIC, or BIC, this bound applies to *any* $f$, not just an $f$ obtained by Max Likelihood. Moreover, (6) is a *finite sample* (i.e. not asymptotic) result, it holds for any value of $n$.

**Theorem 2** *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 $\theta \in (0, 1)$. Denote $\mathcal{D} = \{(x^i, y^i), i = 1 : n\}$ the current training set. Then, with probability w.p. $> 1 - \delta$ over training sets*

$$L_{01}(f) \ \leq \ \hat{L}_{01,\theta}(f) + \tilde{\mathcal{O}}\left(\sqrt{\frac{h}{n\theta^2}}\right) \tag{7}$$

*for any $f \in \mathcal{F}$, where*

$$\hat{L}_{01,\theta}(f) \;=\; \frac{1}{n}|\{i \,|\, y^i f(x^i) \leq \theta\}| \tag{8}$$

This theorem upper bounds the the true loss $L_{01}(f)$ using the number of margin errors for an arbitrary margin $\theta$. Note that for $\theta = 0$ a margin error is also a classification error, and for $\theta > 0$ the number of margin errors is greater or equal to that of classification errors. Hence, the first term of the bound increases with $\theta$, while the second term decreases. So, if most examples can be classified with a large margin (not necessarily 1), then the bound of Theorem 2 can be tighter.

**Structural Risk Minimization (SRM)** is a model selection method. In SRM one selects the model for which the bound on the r.h.s of (6) is minimized. There is one assumption, that the model classes $\mathcal{F}^{1:m}$ from which $f^{1:m}$ are chosen are nested, i.e. $\mathcal{F}^1 \subset \mathcal{F}^2 \subset \ldots \mathcal{F}^m$ with VC dimensions $h^1 < h^2 < \ldots < h^m$. Given $\mathcal{D}$, we fit $f^{1:m}$ on it, and obtain empirical losses $\hat{L}_{01}(f^j)$, $j = 1 : m$. We choose the model for which

$$\hat{L}_{01}(f^j) + \sqrt{\frac{h^j[1 + \log(2n/h^j)] + \log(4/\delta)}{n}}$$

is minimized. While the bound itself is very loose, and will be a bad estimator of the actual $L(f)$, it has been found that this minimization procedure performs well as a model selection criterion.

# 4 Regularization

Regularization is "continuous form" of model selection. We start with a single model class $\mathcal{F}$, which is rich enough so it can fit the data well. Thus, fitting a model $\hat{f} \in \mathcal{F}$ would probably be overfitting the data. We restrict the complexity/effective degrees of freedom/effective number of parameters of $\hat{f}$ by balancing the minimization of the loss $\hat{L}$ with the minimization of the complexity of $\hat{f}$ as measured by a functional $R(f)$.

Hence, we estimate $f$ by minimizing

$$J(f) \;=\; \hat{L}(f) + \lambda R(f), \;\; \lambda \geq 0 \tag{9}$$

In (9) the first term depends on the data, and the second term depends on properties of $f$ alone. This term is called a **regularizer**. One can always

cast the above optimization into a statistical estimation problem. The term that depends on the data is called (formally) the (negative) **log-likelihood**, while the term $\lambda R(f)$ is the (negative) **(log)-prior**. In this paradigm, the minimization in (9) represent a MAP (Maximum A-Posteriori Estimation). The "prior" $R(f)$ is typically favoring "simple" functions (more about this later). Forms of regularization have been in use in statistics for a long time, under the name **shrinkage**.

**Example 2 (Linear Regression with Least Squares cost)** *The (unregularized) problem is*

$$\min_{\beta} \sum_{i=1}^{n} (y^i - \beta^T x^i)^2 \tag{10}$$

*This problem has the closed form solution*

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{11}$$

*with $\mathbf{X}, \mathbf{Y}$ representing respectively the matrix with the inputs $x^i$ as rows, and the vector of corresponding outputs.*

**Ridge regression**

$$\min_{\beta} \sum_{i=1}^{n} (y^i - \beta^T x^i)^2 + \lambda||\beta||^2 \tag{12}$$

*This is a regularized regression, with $R(\beta) = ||\beta||^2$, which favors $\beta$ vectors near 0. This problem has the closed form solution*

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{Y}. \tag{13}$$

**Lasso**

$$\min_{\beta} \sum_{i=1}^{n} (y^i - \beta^T x^i)^2 + \lambda||\beta||_1 \tag{14}$$

*Here, the penalty on $\beta$ is proportional to $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$, the 1-norm of $\beta$. We shall see later that this is a* sparsity inducing *penalty. The Lasso estimator does not have a closed form expression.*

In the upcoming lectures, we shall see other examples of regularization at work.