Advanced Machine Learning from Theory to Practice Lecture 5 Model Selection, Cross Validation and Penalization

F. d'Alche-Buc and E. Le Pennec

Fall 2015

Outline

Experience, Task and Performance measure

- Training data : $\mathcal{D} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ (i.i.d. $\sim \mathbf{P}$)
- Predictor : $f : \mathcal{X} \to \mathcal{Y}$ measurable
- Cost/Loss function : $\ell(Y, f(\mathbf{X}))$ measure how well $f(\mathbf{X})$ "predicts" Y
- Risk :

$$\mathcal{R}(f) = \mathbb{E}\left[\ell(Y, f(\mathbf{X}))\right] = \mathbb{E}_X\left[\mathbb{E}_{Y|\mathbf{X}}\left[\ell(Y, f(\mathbf{X}))\right]\right]$$

- Often $\ell(Y, f(\mathbf{X})) = |f(\mathbf{X}) - Y|^2$ or $\ell(Y, f(\mathbf{X})) = \mathbf{1}_{Y \neq f(\mathbf{X})}$

Goal

- Learn a rule to construct a classifier $\widehat{f} \in \mathcal{F}$ from the training data $\mathcal{D}_n$ s.t. the risk $\mathcal{R}(\widehat{f})$ is small on average or with high probability with respect to $\mathcal{D}_n$.

- The best solution $f^\star$ (which is independent of $\mathcal{D}_n$) is

$$f^\star = \arg\min_{f \in \mathcal{F}} R(f) = \arg\min_{f \in \mathcal{F}} \mathbb{E}\left[\ell(Y, f(\mathbf{X}))\right] = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[\ell(Y, f(\mathbf{x}))\right]\right]$$

Bayes Classifier (explicit solution)

- In binary classification with $0 - 1$ loss :

$$f^\star(\mathbf{X}) = \begin{cases} +1 & \text{if} \quad \mathbb{P}\{Y = +1|\mathbf{X}\} \geq \mathbb{P}\{Y = -1|\mathbf{X}\} \\ & \quad \Leftrightarrow \mathbb{P}\{Y = +1|\mathbf{X}\} \geq 1/2 \\ -1 & \text{otherwise} \end{cases}$$

- In regression with the quadratic loss

$$f^\star(\mathbf{X}) = \mathbb{E}\left[Y|\mathbf{X}\right]$$

Issue : Explicit solution requires to know $\mathbb{E}\left[Y|\mathbf{X}\right]$ for all values of $\mathbf{X}$!

Machine Learning

- Learn a rule to construct a classifier $\widehat{f} \in \mathcal{F}$ from the training data $\mathcal{D}_n$ s.t. the risk $\mathcal{R}(\widehat{f})$ is small on average or with high probability with respect to $\mathcal{D}_n$.

Canonical example : Empirical Risk Minimizer

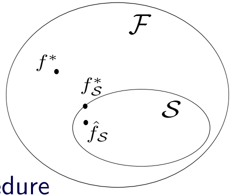
- One restricts f to a subset of functions $\mathcal{S} = \{f_\theta, \theta \in \Theta\}$
- One replaces the minimization of the average loss by the minimization of the empirical loss

$$\widehat{f} = f_{\widehat{\theta}} = \operatorname*{argmin}_{f_\theta, \theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f_\theta(\mathbf{X}_i))$$

- Examples :
  - Linear regression
  - Linear discrimination with

$$\mathcal{S} = \left\{\mathbf{x} \mapsto \mathtt{sign}\{\beta^T \mathbf{x} + \beta_0\} \,/ \beta \in \mathbb{R}^d, \beta_0 \in \mathbb{R}\right\}$$

- General setting :
  - $\mathcal{F} = \{$measurable fonctions $\mathcal{X} \to \mathcal{Y}\}$
  - Best solution : $f^\star = \mathrm{argmin}_{f \in \mathcal{F}} \mathcal{R}(f)$
  - Class $\mathcal{S} \subset \mathcal{F}$ of functions
  - Ideal target in $\mathcal{S}$: $f^\star_{\mathcal{S}} = \mathrm{argmin}_{f \in \mathcal{S}} \mathcal{R}(f)$
  - Estimate in $\mathcal{S}$: $\widehat{f}_{\mathcal{S}}$ obtained with some procedure



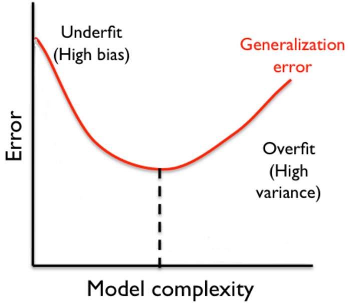
Approximation error and estimation error (Bias/Variance)

$$\mathcal{R}(\widehat{f}_{\mathcal{S}}) - \mathcal{R}(f^\star) = \underbrace{\mathcal{R}(f^\star_{\mathcal{S}}) - \mathcal{R}(f^\star)}_{\text{Approximation error}} + \underbrace{\mathcal{R}(\widehat{f}_{\mathcal{S}}) - \mathcal{R}(f^\star_{\mathcal{S}})}_{\text{Estimation error}}$$

- Approx. error can be large if the model $\mathcal{S}$ is not suitable.
- Estimation error can be large if the model is complex.

Agnostic approach

- No assumption (so far) on the law of $(\mathbf{X}, Y)$.



- Different behavior for different model complexity
- Low complexity model are easily learned but the approximation error ("bias") may be large (Under-fit).
- High complexity model may contains a good ideal target but the estimation error ("variance") can be large (Over-fit)

Bias-variance trade-off $\iff$ avoid overfitting and underfitting

How to find a good function f with a *small* risk

$$R(f) = \mathbb{E}\left[\ell(Y, f(X))\right] \quad ?$$

Canonical approach : $\widehat{f}_{\mathcal{S}} = \operatorname{argmin}_{f \in \mathcal{S}} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(\mathbf{X}_i))$

Problems

- How to choose $\mathcal{S}$?
- How to compute the minimization ?

A Statistical Point of View

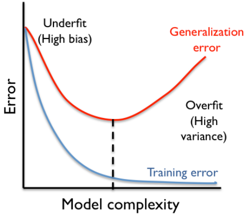
Solution : For $\mathbf{X}$, estimate $Y|\mathbf{X}$ plug this estimate in the Bayes classifier : (Generalized) Linear Models, Kernel methods, k-nn, Naive Bayes, Tree, Bagging...

An Optimization Point of View

Solution : If necessary replace the loss ℓ by an upper bound ℓ' and minimize the empirical loss : SVR, SVM, Neural Network,Tree, Boosting

Models, Complexity and Selection

Over-fitting Issue



Error behaviour

- Learning/training error (error made on the learning/training set) decays when the complexity of the model increases.
- Quite different behavior when the error is computed on new observations (generalization error).

- Overfit for complex models : parameters learned are too specific to the learning set !
- General situation ! (Think of polynomial fit...)
- Need to use an other criterion than the training error !

Two Approaches

- **Cross validation :** Very efficient (and almost always used in practice !) but slightly biased as it target uses only a fraction of the data.
- **Penalization approach :** use empirical loss criterion but penalize it by a term increasing with the complexity of $\mathcal{S}$
$$R_n(\widehat{f_{\mathcal{S}}}) \to R_n(\widehat{f_{\mathcal{S}}}) + \mathrm{pen}(\mathcal{S})$$
and choose the model with the smallest penalized risk.

Which loss to use ?

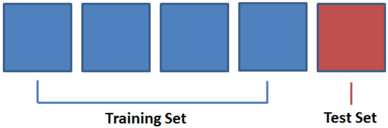
- The loss used in the risk : most natural !
- The loss used to estimate $\hat{\theta}$: penalized estimation !

- **Very simple idea :** use a second learning/verification set to compute a verification error.
- Sufficient to remove the dependency issue !

Cross Validation

- Use $(1-\varepsilon)n$ observations to train and εn to verify !
- Validation for a learning set of size $(1-\varepsilon) \times n$ instead of n !
- Unstable error estimate if εn is too small ?

- Most classical variations :
  - Leave One Out,
  - V-fold cross validation.



Principle

- Split the dataset $\mathcal{D}$ in V sets $\mathcal{D}_v$ of almost equals size.
- For $v \in \{1, .., V\}$:
  - Learn $\hat{f}^{-v}$ from the dataset $\mathcal{D}$ minus the set $\mathcal{D}_v$.
  - Compute the empirical error :

$$R_n^{-v}(\hat{f}^{-v}) = \frac{1}{n_v} \sum_{(\mathbf{X}_i, Y_i) \in \mathcal{D}_v} \ell(Y_i, \hat{f}^{-v}(\mathbf{X}_i))$$

- Compute the average empirical error :

$$R_n^{CV}(\hat{f}) = \frac{1}{|V|} R_n^{-v}(\hat{f}^{-v})$$

Analysis (when n is a multiple of V)

- The $R_n^{-v}(\hat{f}^{-v})$ are identically distributed variable but are not independent !
- Consequence :

$$\mathbb{E}\left[R_n^{CV}(\hat{f})\right] = \mathbb{E}\left[R_n^{-v}(\hat{f}^{-v})\right]$$

$$\mathbb{V}\left[R_n^{CV}(\hat{f})\right] = \frac{1}{V}\mathbb{V}\left[R_n^{-v}(\hat{f}^{-v})\right] + (1-\frac{1}{V})\mathrm{Cov}\left[R_n^{-v}(\hat{f}^{-v}), R_n^{-v'}(\hat{f}^{-v'})\right]$$

- Average risk for a sample of size $(1-\frac{1}{V})n$.
- Variance term much more complex to analyse !
- Fine analysis shows that the larger V the better...

- Accuracy/Speed tradeoff : $V=5$ or $V=10$!

Principle

- The empirical loss computed on an estimator selected in a family according to the data is biased !
- Optimistic estimation of the risk...
- Estimate an upper bound of this optimism for a given family, called the penalty.
- Add it to the empirical loss

- One can also think of the penalty as a way to force the use of *simple* models...

Penalized Loss

- Minimization of

$$\operatorname*{argmin}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f_\theta(\mathbf{X}_i)) + \operatorname{pen}(\theta)$$

  where $\operatorname{pen}(\theta)$ is a penalty.

Penalties

- Upper bound of the optimism of the empirical loss
- Depends on the loss and the framework !

Instantiation

- Penalized Loss for Linear Model
- Structural Risk Minimization

- **Setting** : Gen. linear model = prediction of Y by $h(\mathbf{X}^t\beta)$.

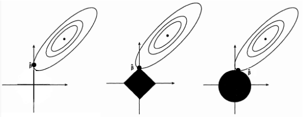
Model coefficients

- Model entirely specified by β.
- Coefficientwise :
  - $\beta_i = 0$ means that the ith covariate is not used.
  - $\beta_i \sim 0$ means that the ith covariate as a *low* influence...

- If some covariates are useless, better use a simpler model...

Submodels

- *Simplify* the model through a constraint on β !
- Examples :
  - Support : Impose that $\beta_i = 0$ for $i \notin I$.
  - Support size : Impose that $\|\beta\|_0 = \sum_{i=1}^d \mathbf{1}_{\beta_i \neq 0} < C$
  - Norm : Impose that $\|\beta\|_p < C$ with $1 \leq p$ (Often $p = 2$ or $p = 1$)



Sparsity

- β is sparse if its number of non-zero coefficients (ℓ_0) is small...
- Easy interpretation in term of dimension/complexity.

Norm Constraint and Sparsity

- Sparsest solution obtained by definition with the ℓ_0 norm.
- No induced sparsity with the ℓ_2 norm...
- Sparsity with the ℓ_1 norm (can even be proved to be the same than with the ℓ_0 norm under some assumptions).
- Geometric explanation.

Constrained Optimization

- Choose a constant C.
- Compute β as

$$\operatorname*{argmin}_{\beta \in \mathbb{R}^d, \|\beta\|_p \leq C} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, h(\langle \mathbf{X}_i, \beta \rangle))$$

Lagrangian Reformulation

- Choose λ and compute β as

$$\operatorname*{argmin}_{\beta \in \mathbb{R}^d} \frac{1}{n} \ell(Y_i, \langle \mathbf{X}_i, \beta \rangle) + \lambda \|\beta\|_p^{p'}$$

  with $p' = p$ except if $p = 0$ where $p' = 1$.
- Easier calibration...

- **Rk :** $\|\beta\|_p$ is not scaling invariant if $p \neq 0$...
- Initial rescaling issue.

Penalized Linear Model

- Minimization of
$$\operatorname*{argmin}_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, h(\beta^t \mathbf{X}_i)) + \mathrm{pen}(\beta)$$
where $\mathrm{pen}(\beta)$ is a (sparsity promoting) penalty
- Variable selection if β is sparse.

Classical Penalties

- AIC : $\mathrm{pen}(\beta) = \lambda \|\beta\|_0$ (non convex / sparsity)
- Ridge : $\mathrm{pen}(\beta) = \lambda \|\beta\|_2^2$ (convex / no sparsity)
- Lasso : $\mathrm{pen}(\beta) = \lambda \|\beta\|_1$ (convex / sparsity)
- Elastic net : $\mathrm{pen}(\beta) = \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$ (convex / sparsity)

- Easy optimization if pen (and the loss) is convex...
- **Need to specify λ !**

Classical Examples

- Penalized Least Squares
- Penalized Logistic Regression
- Penalized Maximum Likelihood
- SVM
- Tree pruning

- Sometimes used even if the parametrization is not linear...

Practical Selection Methodology

- Choose a penalty shape $\widetilde{\text{pen}}$.
- Compute a CV error for a penalty $\lambda\widetilde{\text{pen}}$ for all $\lambda \in \Lambda$.
- Determine $\widehat{\lambda}$ the λ minimizing the CV error.
- Compute the parameters with a penalty $\widehat{\lambda}\widetilde{\text{pen}}$.

Why not using only CV ?

- **If** the penalized likelihood minimization is easy, much cheaper to compute the CV error for all $\lambda \in \Lambda$ than for all possible estimators...
- CV performs best when the set of candidates is not too big (or is structured...)

Penalized $\ell^{0/1}$ loss (Strutural Risk Minimization)

- Minimization of

$$\operatorname*{argmin}_{f_m, m \in \mathcal{M}, f_m \in \mathcal{S}_m} \frac{1}{n} \sum_{i=1}^{n} \ell^{0/1}(Y_i, f_m(\mathbf{X}_i)) + \mathrm{pen}(m)$$

where $\mathrm{pen}(m)$ is a complexity driven penalty...

- No easy optimization here !

Classical Penalties

- Finite class : $\mathrm{pen}(m) = \lambda \sqrt{\frac{\log |\mathcal{M}|}{n}}$
- Finite VC Dimension : $\mathrm{pen}(m) = \lambda \sqrt{\frac{d_{VC}(\mathcal{S}_m) \log\left(\frac{en}{d_{VC}(\mathcal{S}_m)}\right)}{n}}$

- **Need to specify λ !**

Penalized convexified ℓ loss

- Minimization of
$$\operatorname*{argmin}_{f_m, m\in\mathcal{M}, f_m\in\mathcal{S}_m} \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f_m(\mathbf{X}_i)) + \mathrm{pen}(m)$$
where $\mathrm{pen}(m)$ is a complexity driven penalty...

- No easy optimization here !
- **Reuse the previous** $\mathrm{pen}(m)$ **!**
- **Need to specify** λ **!**
- SVM case :
  - $d_{VC} \sim \|\beta\|^2$ which advocates for a penalty in $\lambda\|\beta\|$...
  - A penalty in $\lambda'\|\beta\|^2$ is more convenient numerically and there is a correspondence between the two problems...

Practical Selection Methodology

- Choose a penalty shape $\widetilde{\text{pen}}$.
- Compute a CV error for a penalty $\lambda\widetilde{\text{pen}}$ for all $\lambda \in \Lambda$.
- Determine $\widehat{\lambda}$ the λ minimizing the CV error.
- Compute the final model with a penalty $\widehat{\lambda}\widetilde{\text{pen}}$.

Why not using only CV ?

- **If** the penalized likelihood minimization is easy, much cheaper to compute the CV error for all $\lambda \in \Lambda$ than for all possible estimators...
- CV performs best when the set of candidates is not too big (or is structured...)

ℓ^0 Penalized Empirical Loss Minimization

- Minimization of
$$\operatorname*{argmin}_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1} \ell(Y_i, h(\langle \mathbf{X}_i, \beta \rangle) + \lambda \|\beta\|_0$$
- Equivalent model selection reformulation :
  - For every $I \subset \{1, \ldots, d\}$, compute
$$\hat{\beta}_I = \operatorname*{argmin}_{\beta_J, \beta_{J,i}=0 \forall i \notin I} \frac{1}{n} \sum_{i=1} \ell(Y_i, h(\langle \mathbf{X}_i, \beta \rangle)$$
  - Determine
$$\hat{I} = \operatorname*{argmin}_{i} \frac{1}{n} \sum_{i=1} \ell(Y_i, \langle \mathbf{X}_i, \hat{\beta}_I \rangle) + \lambda |I|$$

- Need to perform those optimization (non convex/non smooth) !
- Need to choose λ (to guaranty good performance) !

Exact Minimization

- Easy optimization for a given support !
- Very different situation for the support...
- Bruteforce exploration of the support = combinatorial problem.
- 2^d models (supports) to be explored !
- Only possible if d is (very) small !

Clever Exploration

- Minimization of the criterion but without an exhaustive exploration of the subsets.
- Generic strategy :
  - Start with a pool of subsets of size P
  - Create a larger pool of size PC by adding and/or removing variables from the previous subset
  - Keep only the best P subset according to the criterion and iterate
- Variations on the size of the subsets, the initial subsets, the rule to add and remove variables, the criterion...
- Forward, Backward, Forward/Backward, Stochastic (Genetic) Algorithm...

Forward strategy

- Start with an empty model
- At each step, create a larger collection by creating models equal to the current one plus any variable not used in the current model (one at a time)
- Modify the current model if the best model within the new collection leads to a reduction of the criterion.

Backward strategy

- Start with the full model.
- At each step, create a larger collection by creating models equal to the current one minus any variable used in the current model (one at a time)
- Modify the current model if the best model within the new collection leads to a reduction of the criterion.

Forward/Backward strategy

- Start with the full model.
- At each step, create a larger collection by creating models equal to the current one plus any variable not used in the current model (one at a time) and to the current one minus any variable used in the current model (one at a time)
- Modify the current model if the best model within the new collection leads to a reduction of the criterion.

- Various Stochastic (Genetic) Algorithm...
- Stability issue...

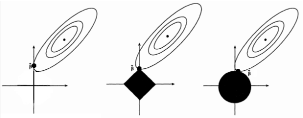
ℓ^1 Penalized Empirical Loss Minimization

- Minimization of
$$\operatorname*{argmin}_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1} \ell(Y_i, h(\langle \mathbf{X}_i, \beta \rangle) + \lambda \|\beta\|_1$$
- Introduced originally as a convexification of the ℓ^0 loss...

- Non smooth but convex function and thus existing fast optimization algorithm.
- Need to choose λ (to guaranty good performance) !

Practical Minimization

ℓ^1 Penalization and Sparsity



Sparsification Properties

- Let
$$L(\beta) = \frac{1}{n}\sum_{i=1} \ell(Y_i, h(\langle \mathbf{X}_i, \beta\rangle) + \lambda\|\beta\|_1$$
- Convex subgradient property : $\widehat{\beta} = \operatorname{argmin} L(\beta) \Leftrightarrow 0 \in \delta L(\widehat{\beta})$:
$$\sum_{i=1}^{n} \mathbf{X}_{i,k} h'(\langle \mathbf{X}_i, \beta\rangle) \frac{d\ell}{dh}(Y_i, h(\langle \mathbf{X}_i, \beta\rangle) \begin{cases} = \lambda & \text{if } \widehat{\beta}_k < 0 \\ \in [-\lambda, \lambda] & \text{if } \widehat{\beta}_k = 0 \\ = -\lambda & \text{if } \widehat{\beta}_k > 0 \end{cases}$$
- More *flexibility* at $\beta_k = 0$...

Penalized ℓ loss (Strutural Risk Minimization)

- Minimization of

$$\operatorname*{argmin}_{f_m, m \in \mathcal{M}, f_m \in \mathcal{S}_m} \frac{1}{n} \sum_{i=1}^{n} \ell^{0/1}(Y_i, f_m(\mathbf{X}_i)) + \mathrm{pen}(m)$$

where $\mathrm{pen}(m)$ is a complexity driven penalty...

- ℓ^0 type penalties...
- No easy optimization here if $\ell = \ell^{0/1}$: full exploration.
- Convex loss with linear classifier : subset exploration.
- SVM relaxation leads to a quadratic convex problem...

Three Examples

- Linear model and unbiased estimate of the risk
- Maximum Likelihood and asymptotic analysis
- Empirical Risk Minimization and concentration

Model and Predictor

- Model : $Y_i = f_0(\mathbf{X}_i) + \sigma\varepsilon_i$ with ε_i i.i.d. such that $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{V}[\varepsilon_i] = 1$.
- Linear predictor : we try to predict y from x by $f_\beta(x) = \sum_{k=1}^{p} \beta_k \varphi_k(x)$

Least Square Approach

$$\widehat{\beta} = \operatorname*{argmin}_{\beta} \sum_{i=1}^{n} \left| Y_i - \sum_{k=1}^{p} \beta_k \varphi_k(\mathbf{X}_i) \right|^2$$

- Geometric interpretation : $(f_{\widehat{\beta}}(\mathbf{X}_i))_{i=1}^n$ is the orthogonal projection $P_V\mathbf{Y}$ of $\mathbf{Y} = (Y_i)_{i=1}^n$ on the space spanned by $V = \{(\varphi_k(\mathbf{X}_i))_{i=1}^n\}_{k=1}^p$.

Prediction error on the same grid

$$
\begin{aligned}
\sum_{i=1}^{n} |y_i' - \sum_{k=1}^{p} f_{\widehat{\beta}}(\mathbf{X}_i)|^2 &= \|(I - P_V)\mathbf{F}_0 + \sigma\varepsilon' - \sigma P_V \varepsilon\|^2 \\
&= \|(I - P_V)\mathbf{F}_0\|^2 + \sigma^2\|\varepsilon'\|^2 + \sigma^2\|P_V\varepsilon\|^2 \\
&\quad + 2\sigma\langle (I - P_V)\mathbf{F}_0, \varepsilon' - P_V\varepsilon\rangle \\
&\quad - 2\sigma^2\langle \varepsilon', P_V\varepsilon\rangle
\end{aligned}
$$

and thus

$$
\mathbb{E}\left[\sum_{i=1}^{n} |y_i' - \sum_{k=1}^{p} f_{\widehat{\beta}}(\mathbf{X}_i)|^2\right] = \|(I - P_V)\mathbf{F}_0\|^2 + n\sigma^2 + p\sigma^2
$$

Empirical error analysis

$$\begin{aligned}
\sum_{i=1} |Y_i - \sum_{k=1}^{p} f_{\widehat{\beta}}(\mathbf{X}_i)|^2 &= \|\mathbf{Y} - P_V \mathbf{Y}\|^2 \\
&= \|(I - P_V)\mathbf{F}_0 + \sigma(I - P_V)\varepsilon\|^2 \\
&= \|(I - P_V)\mathbf{F}_0\|^2 + \sigma^2 \|(I - P_V)\varepsilon\|^2 \\
&\qquad + 2\sigma \langle (I - P_V)\mathbf{F}_0, (I - P_V)\varepsilon \rangle
\end{aligned}$$

and thus

$$\mathbb{E}\left[\sum_{i=1} |Y_i - \sum_{k=1}^{p} f_{\widehat{\beta}}(\mathbf{X}_i)|^2\right] = \|(I - P_V)\mathbf{F}_0\|^2 + (n-p)\sigma^2$$

Relationship between the two expectations

$$\mathbb{E}\left[\sum_{i=1}^{n} |y_i' - \sum_{k=1}^{p} f_{\widehat{\beta}}(\mathbf{X}_i)|^2\right] = \mathbb{E}\left[\sum_{i=1} |\, Y_i - \sum_{k=1}^{p} f_{\widehat{\beta}}(\mathbf{X}_i)|^2\right] + 2p\sigma^2$$

- Unbiased estimation heuristic : add a penalty of $2p\sigma^2$ to the empirical error to correct the bias...

Likelihood and Contrast

- Likelihood :

$$L_n(\theta) = \sum_{i=1}^{n} \log p_\theta(Y_i|\mathbf{X}_i)$$

- True contrast :

$$L(\theta) = \mathbb{E}_{(X,Y)}\left[\log p_\theta(Y|X)\right]$$

- Maximum Likelihood and target :

$$\widehat{\theta} = \arg\max L_{(\mathbf{X}_i,Y_i)}(\theta) \quad \text{and} \quad \widetilde{\theta} = \arg\max L(\theta)$$

- Taylor expansion around $\widetilde{\theta}$

$$L_n(\theta) \sim L_n(\widetilde{\theta}) + \nabla L_n(\widetilde{\theta})^t(\theta - \widetilde{\theta}) + \frac{1}{2}(\theta - \widetilde{\theta})^t HL_n(\widetilde{\theta})(\theta - \widetilde{\theta})$$

$$L(\theta) \sim L(\widetilde{\theta}) + \frac{1}{2}(\theta - \widetilde{\theta}) HL_n(\widetilde{\theta})(\theta - \widetilde{\theta})$$

- We deduce

$$\widehat{\theta} \sim \widetilde{\theta} - (HL_n(\widetilde{\theta}))^{-1}\nabla L_n(\widetilde{\theta})$$

and thus

$$L_n(\widehat{\theta}) \sim L_n(\widetilde{\theta}) - \frac{1}{2}\nabla L_n(\widetilde{\theta})(HL_n(\widetilde{\theta}))^{-1}\nabla L_n(\widetilde{\theta})$$

$$L(\widehat{\theta}) \sim L(\widetilde{\theta}) + \frac{1}{2}\nabla L_n(\widetilde{\theta})(HL_n(\widetilde{\theta}))^{-1} HL(\widetilde{\theta})(HL_n(\widetilde{\theta}))^{-1}\nabla L_n(\widetilde{\theta})$$

- Thus
$$\begin{aligned}L(\widehat{\theta}) - L_n(\theta) \sim L(\widetilde{\theta}) - L_n(\widetilde{\theta}) \\ &+ \frac{1}{2}\nabla L_n(\widetilde{\theta})^t (HL_n(\widetilde{\theta}))^{-1} HL(\widetilde{\theta}) (HL_n(\widetilde{\theta}))^{-1} \nabla L_n(\widetilde{\theta}) \\ &+ \frac{1}{2}\nabla L_n(\widetilde{\theta})^t (HL_n(\widetilde{\theta}))^{-1} \nabla L_n(\widetilde{\theta})\end{aligned}$$
- As $HL_n(\widetilde{\theta})$ tends to $HL(\widetilde{\theta})$, we have
$$L(\widehat{\theta}) - L_n(\theta) \sim L(\widetilde{\theta}) - L_n(\widetilde{\theta}) + \nabla L_n(\widetilde{\theta})^t (HL(\widetilde{\theta}))^{-1} \nabla L_n(\widetilde{\theta}))$$
- Now, by the CLT,
$$\sqrt{n}\nabla L_n(\widetilde{(\theta)}) \to Z \sim \mathcal{N}(0, J(\widetilde{\theta}))$$
with $J(\theta) = \mathbb{V}\left[\nabla \log p_\theta(X|Y)\right]$ and thus
$$L(\widehat{\theta}) - L_n(\theta) \sim L(\widetilde{\theta}) - L_n(\widetilde{\theta}) + \frac{1}{n} Z^t HL(\widetilde{\theta})^{-1} Z.$$
- Taking the expectation leads to
$$\mathbb{E}\left[L(\widehat{\theta}) - L_n(\widehat{\theta})\right] \sim \frac{1}{n}\mathrm{Tr}(HL(\widetilde{\theta})^{-1} J(\widetilde{\theta}))$$

- Now,

$$\begin{aligned}
HL(\widetilde{\theta})_{i,j} &= \mathbb{E}\left[\frac{\partial^2}{d\theta_i d\theta_j}(\log p_{\widetilde{\theta}}(X,Y))\right] \\
&= \mathbb{E}\left[-\frac{\frac{\partial}{d\theta_j}p_{\widetilde{\theta}}(X,Y) \times \frac{\partial}{d\theta_j}p_{\widetilde{\theta}}(X,Y)}{p^2_{\widetilde{\theta}}(X,Y)}\right] + \mathbb{E}\left[\frac{\frac{\partial^2}{d\theta_j d\theta_i}p_{\widetilde{\theta}}(X,Y)}{p_{\widetilde{\theta}}(X,Y)}\right] \\
&= -\mathbb{E}\left[\left(\frac{\partial}{d\theta_i}\log p_{\widetilde{\theta}}(X,Y)\right)\left(\frac{\partial}{d\theta_j}\log p_{\widetilde{\theta}}(X,Y)\right)\right] + \Delta(\widetilde{\theta}) \\
&= J(\widetilde{\theta}) + \Delta(\widetilde{\theta})
\end{aligned}$$

with $\Delta(\widetilde{\theta}) = \mathbb{E}\left[\frac{\frac{\partial^2}{d\theta_j d\theta_i}p_{\widetilde{\theta}}(X,Y)}{p_{\widetilde{\theta}}(X,Y)}\right]$

Asymptotic Control

- We have thus

$$\mathbb{E}\left[L(\widehat{\theta}) - L_n(\widehat{\theta})\right] \sim -\frac{p}{n} - \frac{1}{n}\mathrm{Tr}(HL(\widetilde{\theta})^{-1}\Delta(\widetilde{\theta}))$$

- Note that if $p_{\widetilde{\theta}}$ is the true law then

$$\Delta(\widetilde{\theta}) = \mathbb{E}\left[\frac{\frac{\partial^2}{d\theta_j d\theta_i} p_{\widetilde{\theta}}(X,Y)}{p_{\widetilde{\theta}}(X,Y)}\right]$$

$$= \int \frac{\partial^2}{d\theta_j d\theta_i} p_{\widetilde{\theta}}(x,y) dxdy = 0$$

and we obtain a **bias of** p/n! (Usual AIC)

Bayesian Approach

- Bayesian Approach :

$$
\begin{aligned}
\log \mathbb{P}\left\{\mathcal{M}|Y\right\} &= \log \int \mathbb{P}\left\{\mathcal{M},\theta|Y\right\} d\theta \\
&= \log \int \frac{\mathbb{P}\left\{Y|\theta,\mathcal{M}\right\}\mathbb{P}\left\{\theta|\mathcal{M}\mathbb{P}\left\{M\right\}\right\}}{\mathbb{P}\left\{Y\right\}} d\theta \\
&= \log \int e^{n(L_n(\theta)+\frac{1}{n}\log \mathbb{P}\{\theta|\mathcal{M}\})} d\theta \\
&\quad + \log \mathbb{P}\left\{M\right\} - \log \mathbb{P}\left\{Y\right\}
\end{aligned}
$$

- Using a Taylor expansion around $\widehat{\theta}$ yields
  $L_n(\theta) \sim L_n(\widehat{\theta}) - \frac{1}{2}(\theta - \widehat{\theta})^t(-HL_n(\widehat{\theta}))(\theta - \widehat{\theta})$
- We deduce that
  $$\log \int e^{n(L_n(\theta)+\frac{1}{n}\log \mathbb{P}\{\theta|\mathcal{M}\})} d\theta \sim nL_n(\widehat{\theta}) \\ + \log \int e^{-\frac{1}{2}(\theta-\widehat{\theta})^t(-nHL_n(\widehat{\theta}))(\theta-\widehat{\theta})+\log \mathbb{P}\{\theta|\mathcal{M}\}} d\theta$$
- If we assume the prior is flat around $\widehat{\theta}$, we obtain
  $$\log \int e^{n(L_n(\widehat{\theta})+\frac{1}{n}\log \mathbb{P}\{\theta|\mathcal{M}\})} d\theta \sim nL_n(\theta) + \log \mathbb{P}\left\{\widehat{\theta})|\mathcal{M}\right\} \\ + \frac{d}{2}\log\frac{2\pi}{n} - \frac{1}{2}\log\det(-HL_n(\widehat{\theta}))$$

- Hence

$$\log \mathbb{P}\left\{\mathcal{M}|Y\right\} \sim n\left(L_n(\widehat{\theta}) - \frac{\log n - \log 2\pi}{2}\frac{d}{n}\right)$$
$$- \frac{1}{2}\log \det(-HL_n(\widehat{\theta}))$$
$$+ \log \mathbb{P}\left\{\widehat{\theta})|\mathcal{M}\right\} + \log \mathbb{P}\left\{M\right\} - \log \mathbb{P}\left\{Y\right\}$$
$$\sim n\left(L_n(\widehat{\theta}) - \frac{\log n}{2}\frac{d}{n}\right)$$

Bayesian Information Criterion

$$\log \mathbb{P}\left\{\mathcal{M}|Y\right\} \sim n\left(L_n(\widehat{\theta}) - \frac{\log n}{2}\frac{d}{n}\right)$$
$$-\log \mathbb{P}\left\{\mathcal{M}|Y\right\} \sim n\left(-L_n(\widehat{\theta}) + \frac{\log n}{2}\frac{d}{n}\right)$$

- Theoretical control of the random (error estimation) term :

$$\mathcal{R}(\widehat{f}_{\mathcal{S}}) - \mathcal{R}(f^*_{\mathcal{S}})$$

Probably Almost Correct Analysis

- **Theoretical guarantee** that with probablity larger than $1 - \delta$,

$$\mathbb{P}\left\{\mathcal{R}(\widehat{f}) - \mathcal{R}(f^\star_{\mathcal{S}}) \leq \varepsilon_{\mathcal{S}}(\delta)\right\} \geq 1 - \delta$$

  for a suitable $\varepsilon_{\mathcal{S}}(\delta) \geq 0$.
- Implies :
  - $\mathbb{P}\left\{\mathcal{R}(\widehat{f}) - \mathcal{R}(f^\star) \leq \mathcal{R}(f^\star_{\mathcal{S}}) - \mathcal{R}(f^\star) + \varepsilon_{\mathcal{S}}(\delta)\right\} \geq 1 - \delta$
  - $\mathbb{E}\left[\mathcal{R}(\widehat{f}) - \mathcal{R}(f^\star_{\mathcal{S}}\right] \leq \int_0^{+\infty} \varepsilon^-_{\mathcal{S}}(t) dt$

- The result should hold without any assumption on the law **P** !

- By construction :

$$\begin{aligned}\mathcal{R}(\widehat{f}) - \mathcal{R}(f^\star_{\mathcal{S}}) &= \mathcal{R}(\widehat{f}) - \mathcal{R}_n(\widehat{f}) + \mathcal{R}_n(\widehat{f}) - \mathcal{R}_n(f^\star_{\mathcal{S}}) + \mathcal{R}_n(f^\star_{\mathcal{S}}) - \mathcal{R}(f^\star_{\mathcal{S}})\\ &\leq \mathcal{R}(\widehat{f}) - \mathcal{R}_n(\widehat{f}) + \mathcal{R}_n(f^\star_{\mathcal{S}}) - \mathcal{R}(f^\star_{\mathcal{S}})\\ &\leq \left(\mathcal{R}(\widehat{f}) - \mathcal{R}(f^\star_{\mathcal{S}})\right) - \left(\mathcal{R}_n(\widehat{f}) - \mathcal{R}_n(f^\star_{\mathcal{S}})\right)\end{aligned}$$

Four possible upperbounds

- $\mathcal{R}(\widehat{f}) - \mathcal{R}(f^\star_{\mathcal{S}}) \leq \sup_{f\in\mathcal{S}} \left((\mathcal{R}(f) - \mathcal{R}(f^\star_{\mathcal{S}})) - (\mathcal{R}_n(f) - \mathcal{R}_n(f^\star_{\mathcal{S}}))\right)$
- $\mathcal{R}(\widehat{f}) - \mathcal{R}(f^\star_{\mathcal{S}}) \leq \sup_{f\in\mathcal{S}} (\mathcal{R}(f) - \mathcal{R}_n(f)) + (\mathcal{R}_n(f^\star_{\mathcal{S}}) - \mathcal{R}(f^\star_{\mathcal{S}}))$
- $\mathcal{R}(\widehat{f}) - \mathcal{R}(f^\star_{\mathcal{S}}) \leq \sup_{f\in\mathcal{S}} (\mathcal{R}(f) - \mathcal{R}_n(f)) + \sup_{f\in\mathcal{S}} (\mathcal{R}_n(f) - \mathcal{R}(f))$
- $\mathcal{R}(\widehat{f}) - \mathcal{R}(f^\star_{\mathcal{S}}) \leq 2\sup_{f\in\mathcal{S}} |\mathcal{R}(f) - \mathcal{R}_n(f)|$

- Supremum of centered random variables !
- **Key :** Concentration of each variable...

- By construction, for any $f' \in \mathcal{S}$,
$$\mathcal{R}(f') = \mathcal{R}_n(f') + \big(\mathcal{R}(f') - \mathcal{R}_n(f')\big)$$

A uniform upper bound for the error

- Simultaneously $\forall f' \in \mathcal{S}$,
$$\mathcal{R}(f') \leq \mathcal{R}_n(f') + \sup_{f \in \mathcal{S}} (\mathcal{R}(f) - \mathcal{R}_n(f))$$

- Supremum of centered random variables !
- **Key :** Concentration of each variable...
- Can be interpreted as a justification of the ERM !

- Empirical loss :

$$\mathcal{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell^{0/1}(Y_i, f(\mathbf{X}_i))$$

Properties

- $\ell^{0/1}(Y_i, f(\mathbf{X}_i))$ are i.i.d. random variables in $[0, 1]$.

Concentration

$$\mathbb{P}\left\{\mathcal{R}(f) - \mathcal{R}_n(f) \leq \varepsilon\right\} \geq 1 - e^{-2n\varepsilon^2}$$

$$\mathbb{P}\left\{\mathcal{R}_n(f) - \mathcal{R}(f) \leq \varepsilon\right\} \geq 1 - e^{-2n\varepsilon^2}$$

$$\mathbb{P}\left\{|\mathcal{R}_n(f) - \mathcal{R}(f)| \leq \varepsilon\right\} \geq 1 - 2e^{-2n\varepsilon^2}$$

- Concentration of sum of bounded independent variables !
- Hoeffding theorem.

Concentration

- If $\mathcal{S}$ is finite of cardinality $|\mathcal{S}|$,

$$\mathbb{P}\left\{\sup_f \left(\mathcal{R}_n(f) - \mathcal{R}(f)\right) \leq \sqrt{\frac{\log|\mathcal{S}| + \log(1/\delta)}{2n}}\right\} \geq 1 - \delta$$

$$\mathbb{P}\left\{\sup_f \left|\mathcal{R}_n(f) - \mathcal{R}(f)\right| \leq \sqrt{\frac{\log|\mathcal{S}| + \log(1/\delta)}{2n}}\right\} \geq 1 - 2\delta$$

- Control of the supremum by a quantity depending on the cardinality and the probability parameter δ.
- Simple combination of Hoeffding and a union bound.

PAC Bounds

- If $\mathcal{S}$ is finite of cardinality $|\mathcal{S}|$, with proba greater than $1-2\delta$
$$\mathcal{R}(\widehat{f}) - \mathcal{R}(f^\star_{\mathcal{S}}) \leq \sqrt{\frac{\log|\mathcal{S}|}{2n}} + \sqrt{\frac{2\log(1/\delta)}{n}}$$
- If $\mathcal{S}$ is finite of cardinality $|\mathcal{S}|$, with proba greater than $1-\delta$, simultaneously $\forall f' \in \mathcal{S}$,
$$\mathcal{R}(f') \leq \mathcal{R}_n(f') + \sqrt{\frac{\log|\mathcal{S}|}{2n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

- Risk increases with the cardinality of $\mathcal{S}$.
- Similar issue in cross-validation !
- No direct extension for an infinite $\mathcal{S}$...

PAC Bounds

- If $\mathcal{S}$ is of VC dimension d_{VC} then if $n > d_{VC}$
- With probability greater than $1 - 2\delta$,

$$\mathcal{R}(\widehat{f}) - \mathcal{R}(f^\star_{\mathcal{S}}) \leq \sqrt{\frac{8 d_{VC} \log\left(\frac{en}{d_{VC}}\right)}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}}$$

- With probability greater than $1 - \delta$, simultaneously $\forall f' \in \mathcal{S}$,

$$\mathcal{R}(f') \leq \mathcal{R}_n(f') + \sqrt{\frac{8 d_{VC} \log\left(\frac{en}{d_{VC}}\right)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

- **Rk :** If $d_{VC} = +\infty$ no uniform PAC bounds exists !

- Assume we have a countable collection of set $(\mathcal{S}_m)_{m\in\mathcal{M}}$ and let π_m be such that $\sum_{m\in\mathcal{M}} \pi_m = 1$.

Non Uniform Risk Bound

- With probability $1-\delta$, simultaneously for all $m \in \mathcal{M}$ and all $f \in \mathcal{S}_m$,

$$\mathcal{R}(f) \leq \mathcal{R}_n(f) + \sqrt{\frac{8 d_{VC}(\mathcal{S}_m) \log\left(\frac{en}{d_{VC}(\mathcal{S}_m)}\right)}{n}} + \sqrt{\frac{\log \pi_m}{2n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

Structural Risk Minimization

- Choose $\hat{f}$ as the minimizer over $m \in \mathcal{M}$ and $f \in \mathcal{S}_m$ of

$$\mathcal{R}_n(f) + \sqrt{\frac{8 d_{VC}(\mathcal{S}_m) \log\left(\frac{en}{d_{VC}(\mathcal{S}_m)}\right)}{n}} + \sqrt{\frac{\log \pi_m}{2n}}$$

- Mimics the minimization of the integrated risk !

PAC Bound

- If $\hat{f}$ is the SRM minimizer then with probability $1 - 2\delta$,

$$\mathcal{R}(\widehat{f}) \leq \inf_{m \in \mathcal{M}} \inf_{f \in \mathcal{S}_m} \left(\mathcal{R}(f) + \sqrt{\frac{8 d_{VC}(\mathcal{S}_m) \log\left(\frac{en}{d_{VC}(\mathcal{S}_m)}\right)}{n}} + \sqrt{\frac{\log \pi_m}{2n}} \right) + \sqrt{\frac{2\log(1/\delta)}{n}}$$

- The SRM minimizer balances the risk $\mathcal{R}(f)$ and the upper bound on the estimation error $\sqrt{\frac{8 d_{VC}(\mathcal{S}_m) \log\left(\frac{en}{d_{VC}(\mathcal{S}_m)}\right)}{n}} + \sqrt{\frac{\log \pi_m}{2n}}$.