Statisitcs

LLN

$$\lim_{n\to\infty}\left|\frac{1}{n}\sum_{i=1}^{n} x_i - \mathbb{E}\left[x\right]\right| = 0$$

Unbiasedness

$$\text{Bias}_\theta = \mathbb{E}\left[\hat\theta\right] - \theta = \mathbb{E}\left[\hat\theta - \theta\right] = 0$$

Consistency

$$\lim_{n\to\infty}\mathbf{P}\left(\left|\widehat{X} - \frac{1}{n}\sum_{i=1}^{n} X_i\right| \le \epsilon\right) = 1$$

Properties of Gaussians

$$X \sim \mathcal{N}(\mu,\, \sigma^2) \implies Y = aX + b \sim \mathcal{N}(a\mu + b,\, a^2\sigma^2)$$
$$X \sim \mathcal{N}(\mu_x,\, \sigma_x^2), \qquad Y \sim \mathcal{N}(\mu_y,\, \sigma_y^2), \qquad \implies \quad Z = X + Y \sim \mathcal{N}(\mu_x + \mu_y,\, \sigma_x^2 + \sigma_y^2)$$

Information Theory

Information

$$\mathbf{I}\left(S\right) = \log_2\frac{1}{\mathbf{P}\left(S\right)}$$

Entropy

$$\begin{aligned}\mathbf{H}\left[S\right] &= \sum_{s\in\{S\}}\mathbf{P}\left(S{=}s\right)\cdot\mathbf{I}\left(S{=}s\right) = \sum_{s\in\{S\}}\mathbf{P}\left(S{=}s\right)\log_2\frac{1}{\mathbf{P}\left(S{=}s\right)}\\ &= \mathbb{E}\left[\log_2 {}^{1}/_{\mathbf{P}\left(S=s\right)}\right] \qquad = -\mathbb{E}\left[\log_2\mathbf{P}\left(S{=}s\right)\right]\\ &= -\sum_{s\in\{S\}}\mathbf{P}\left(S{=}s\right)\log_2\mathbf{P}\left(S{=}s\right)\end{aligned}$$

Information Gain

$$\mathbf{I}\left(Y,\, X_i\right) = \mathbf{H}\left[Y\right] - \mathbf{H}\left[Y \,|\, X_i\right]$$

Learning Theory

PAC Learning

$$\mathbf{P}\left(\left|\hat\theta - \theta^*\right| \le \epsilon\right) \ge 1 - \delta$$
$$\mathbf{P}\left(\left|\hat\theta - \theta^*\right| > \epsilon\right) < \delta$$

Decision Theory

Risk

$$\text{Risk}(f) = R(f) := \mathbb{E}_{(X,Y)}[\text{loss}(Y,\, f(X)]$$

Bayes Risk

$$R(f^*) \le R(f),\ \forall f$$

Bayes Optimal Rule

$$f^*(P) = \arg\min_f \mathbb{E}_{(X,Y)\sim P}[\text{loss}(Y,\, f(X))]$$

Emperical Risk Minimization

$$\hat f_n = \arg\min_f \frac{1}{n}\sum_{i=1}^{n}[\text{loss}(Y_i,\, f(x_i))] \xrightarrow[n\to\infty]{\text{LNN}} \arg\min_f \mathbb{E}_{(X,Y)}[\text{loss}(Y,\, f(X))]$$

Risk

True Risk

$$R(f) := \mathbb{E}_{(X,Y)}[\ell(f(X),\, Y)]$$

Emperical Risk

$$\widehat{R}_{\mathcal{D}}(f) := \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}}[\ell(f(X_i),\, Y_i)]$$

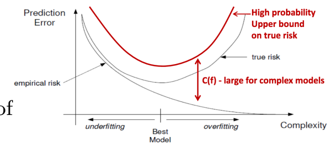
Excess Risk

$$\mathbb{E}\left[R(\hat f_n)\right] - R(f^*)$$

Structural Risk

$$\left|R(f) - \widehat{R}_n(f)\right| \le C(f) \qquad \forall f \in \mathcal{F}$$
$$R(f) \le \widehat{R}_n + C(f), \qquad \forall f \in \mathcal{F}$$

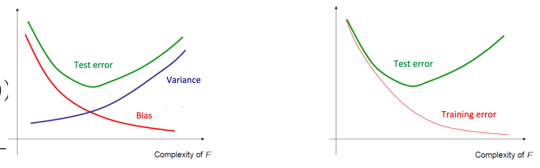
Use $\widehat{R}_n(\hat f_n) + C(\hat f_n)$ as a *pessimistic* estimate of true risk.



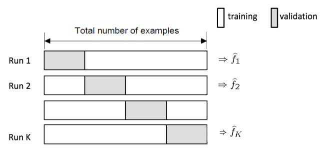
Risk Estimation

Bias–Variance Tradeoff

$$\text{Bias} = \mathbb{E}\left[f(X)\right] - f^*(X)$$
$$\text{Variance} = \mathbb{E}\left[(f(X) - \mathbb{E}\left[f(X)\right.\right.$$
$$\mathbb{E}\left[(f(X) - f^*(X))^2\right] = \text{Bias}^2 +$$

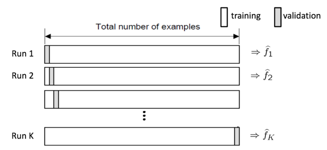


K–Fold Cross–Validation



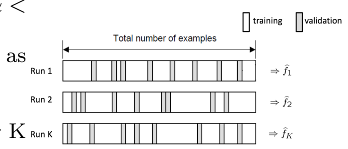
LOO Cross–Validation

Special Case of K–Fold with $K = n$ partitions.



Random Subsampling

Randomly subsample a fixed fraction αn $(0 < \alpha < 1)$ of the dataset for validation.
Form hold–out predictor with remaining data as training data.
Repeat K times.
Final Predictor is average / majority vote over K hold–out estimates



Hold–out Method

$$\mathcal{D} = \left\{(X_i,\, Y_i)\right\}_{i=1}^{n}$$

1. Split into two sets

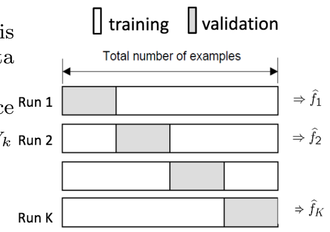
$$\underbrace{\mathcal{D}_T = \left\{(X_i,\, Y_i)\right\}_{i=1}^{m}}_{\text{training set}} \qquad \underbrace{\mathcal{D}_V = \left\{(X_i,\, Y_i)\right\}_{i=m+1}^{n}}_{\text{holdout set}}$$

2. use $\mathcal{D}_T$ to train a predictor $\hat f_{\mathcal{D}_T}$
3. use $\mathcal{D}_V$ to evaluate the predictor $\widehat{R}_{\mathcal{D}_V}(\hat f_{\mathcal{D}_T})$

Estimating True Risk

Estimate the error of a predictor on n data points.
If K is large (close to n), bias of error estimate is small since each training set has close to n data points.
However, variance of error estiamtes is high since each validation set has fewer data points and $\widehat{R}_{V_k}$ might deviate a lot from the mean.

$$\text{Error estimate} = \frac{1}{K}\sum_{k=1}^{K}\widehat{R}_{V_k}(\hat f_{T_k})$$



Risk Minimization

Emperical Risk Minimization

$$\widehat{f_n} = \arg\min_f \frac{1}{n}\sum_{i=1} n[\text{loss}(Y_i,\, f(x_i))]$$

Structural Risk Minimization

$$\hat f_n = \arg\min_{f\in\mathcal{F}}[\widehat{R}_n(f) + \lambda C(f)]$$

Overfitting

Regularization

Complexity Regularization

$$\hat f_n = \arg\min_{f\in\mathcal{F}}\{\widehat{R}_n(f) + C(f)\}$$

Information Criteria

AIC (Akiake IC) $\quad C(f) = \#parameters$
Allows # parameters to be infinite as # training data n becomes large

BIC (Bayesian IC) $\quad C(f) = \#parameters * \log n$
Penalizes complex models more heavily – limits complexity of models as # training data n becomes large

Model Selection

- define a finite set of model classes
- estimate true risk for each model class
- select model class with lowest estimated true risk

Model classes $\{\mathcal{F}_\lambda\}$ of increasing complexity $\mathcal{F}_1 < \mathcal{F}_2 < \ldots$

$$\min_\lambda \min_{f\in\mathcal{F}_\lambda} J(f,\, \lambda)$$

- given λ estimate $\hat f_\lambda$ using *emperical / structural / complexity regularized* risk minimization
- select λ for which $\hat f_\lambda$ has minimum true risk estimated using *cross–validation / hold–out / information criteria*

Generalization Error

Terms

estimated predictor $:= \hat f_n$

optimal predictor $:= f^*$

risk of estimated predictor $:= R(\hat f_n)$

expected risk $:= \mathbb{E}\left[R(\hat f_n)\right]$

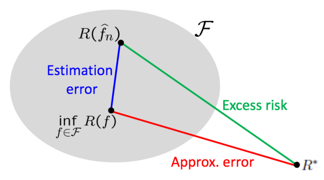
risk of optimal predictor $:= R(f^*)$

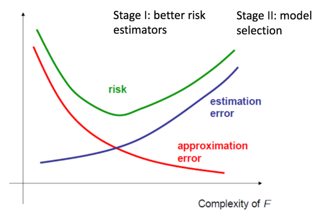
excess risk $:= \mathbb{E}\left[R(\hat f_n)\right] - R(f^*)$

True Risk Decomposition

$$\mathbb{E}\left[R(\hat{f}_n)\right] - R^* = \underbrace{\left(\mathbb{E}\left[R(\hat{f}_n)\right] - \inf_{f\in\mathcal{F}} R(f)\right)}_{\text{estimation error}} + \underbrace{\left(\inf_{f\in\mathcal{F}} R(f) - R^*\right)}_{\text{approximation error}}$$

Estimation Error : due to randomness of training data
Approximation Error : due to restriction of model class





Regression

Linear Regression

Ridge Regression

$$\hat{\theta}_{\mathrm{MAP}} \arg\min_{\theta} \sum_{i=1}^{n} (Y_i - X_i\theta)^2 + \lambda\|\theta\|_2^2$$

Lasso Regression

$$\hat{\theta}_{\mathrm{MAP}} \arg\min_{\theta} \sum_{i=1}^{n} (Y_i - X_i\theta)^2 + \lambda\|\theta\|_1$$

Polynomial Regression

Classification

Logistic Regression

Naive Bayes

Boosting

Decision Trees

$$\begin{aligned}
\arg\max_{X_i} \left[\mathbf{H}\left[Y\right] - \mathbf{H}\left[Y \,\middle|\, X_i\right] \right] &= \arg\min_{X_i} \mathbf{H}\left[Y \,\middle|\, X_i\right] \\
&= \arg\min_{X_i} \sum_{x \in \{X_i\}} \left[\mathbf{P}\left(X_i = x\right) \mathbf{H}\left[Y \,\middle|\, X_i = x\right] \right] \\
&= \arg\min_{X_i} - \sum_{x \in \{X_i\}} \left[\mathbf{P}\left(X_i = x\right) \sum_{y \in \{Y\}} \left[\mathbf{P}\left(Y = \,\middle|\, X_i = x\right) \log_2 \mathbf{P}\left(Y = y \,\middle|\, X_i = x\right) \right] \right]
\end{aligned}$$

Support Vector Machines

Deep Learning

Common Activation Functions

Backpropagation

Gradient Descent

Perceptron

MLP

CNN

RNN

LSTM

Clustering

KNN

Kernel Regression

Kernel Trick