Statistics

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

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

Information Theory

Information

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

Information Gain

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

Entropy

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

Learning Theory

PAC Learning

$$\mathbf{P}\left(\left|\hat{\theta}-\theta^*\right|\le\epsilon\right)\ge1-\delta\qquad\left|\qquad\mathbf{P}\left(\left|\hat{\theta}-\theta^*\right|>\epsilon\right)<\delta\right.$$

Decision Theory

Risk

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

Bayes Risk

$$R(f^*)\le R(f),\ \forall\,f\in\mathcal{F}$$

Bayes Optimal Rule

$$f^*(P)=\arg\min_f\mathbb{E}_{(X,Y)\sim P}\left[\ell\big(Y,\,f(X)\big)\right]$$

Emperical Risk Minimization

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

Risk

True Risk

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

Emperical Risk

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

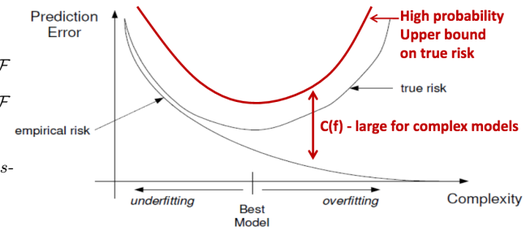
Excess Risk

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

Structural Risk

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

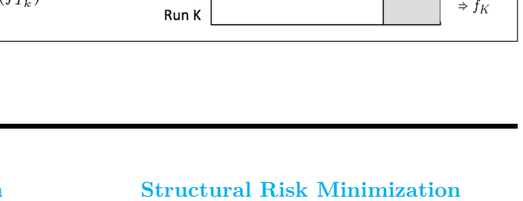
Use $\widehat{R}_n(\hat{f}_n)+C(\hat{f}_n)$ as a *pessimistic* estimate of true risk.



Risk Estimation

Bias–Variance Tradeoff

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



K–Fold CV

LOO CV

Random Subsampling

Hold–out Method

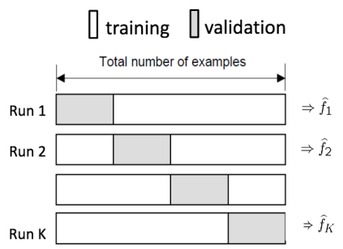
1. split into two sets
2. use $\mathcal{D}_T$ to train a predictor $\hat{f}_{\mathcal{D}_T}$
3. use $\mathcal{D}_V$ to evaluate the predictor, $\widehat{R}_{\mathcal{D}_V}(\hat{f}_{\mathcal{D}_T})$

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

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

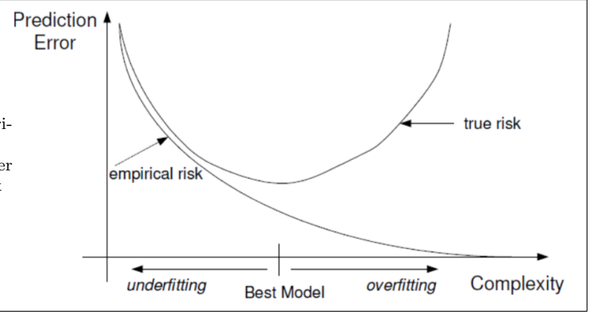
$$\widehat{f_n}=\arg\min_f\frac{1}{n}\sum_{i=1}^{n}\left[\ell\Big(Y_i,\,f(x_i)\Big)\right]$$

Structural Risk Minimization

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

Overfitting

discrepency between emperical risk and true risk, so emperical risk is no longer a good indicator of true risk



Regularization

Complexity Regularization

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

$C(f)$: bound on deviation from true risk
λ : regularization penality

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 $\mathcal{F}_1<\mathcal{F}_2<\dots$

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

- 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

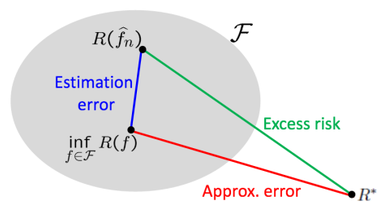
Terms

estimated predictor $:=\hat{f}_n$	optimal predictor $:=f^*$
risk of estimated predictor $:=R(\hat{f}_n)$	risk of optimal predictor $:=R(f^*)$
expected risk $:=\mathbb{E}\left[R(\hat{f}_n)\right]$	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

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

Lasso Regression

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

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