Statistical Learning

Chengxin Gong, Peking University
wqgcx.github.io

2022 年 4 月 6 日

目录

1 Introduction

- What's ML?

  Computational and statistical methods using experience/data to improve performance in various learning tasks, make accurate prediction or estimate inference. Related: SL, DL, AI, DM, Big Data, Data Science.

- References

  1) The Elements of Statistical Learning: Data Mining, Inference, and Prediction.
  2) Foundations of Machine Learning.

- Learning tasks

  $$\text{How data are labeled}\begin{cases}\text{Supervised: class \& regression}\\ \text{Unsupervised: clustering, dim reduct, GM, ranking}\\ \text{Semisupervised: labeled + unlabeled}\end{cases}$$

  $$\text{How data arrive}\begin{cases}\text{Batch/Offline}\\ \text{Online}\end{cases}$$

- Terminology

  Sample: set of sample points(examples/instances)
  i.i.d. $X = (x_1, \cdots, x_n) \sim D^n$. Repeated patterns.
  Relaxed: transfer learning & federated learning.
  Usually, $x_i \in R^p$(feature/statistics), $y_i \in \{0, 1\}$ (labels/outcomes).
  Hypothesis/parameters: functions mapping x_i to y_i, $y = f(x, \theta) + \epsilon$
  Hyperparameters/tuning parameters: free parameters of learning algorithms.
  Simplified: Data → Algorithm(Hyperparameters) → Hypothesis → Predict.
  Realistic: put samples into 3 parts: training, validation and test. Using traning samples to train algorithm and produce hypothesis, but hypothesis depends on hyperparameters, so we use validation samples to choose the best hyperparameters.

- Loss function: $\mathcal{L} : y \times y' \to R_2$, e.g. $\mathcal{L}(y, y') = I(y \neq y')$ or $(y - y')^2$.

- Performance guarantees: generalization/predict, stability/robustness, explainablity/interpretability, computability/scalability

- Occam's razor: Among hypotheses equally consistent(taking noise into account) with data, simpler is better.

- No-free-lunch thm: Learning is impossible without prior knowledge.

  Proof: Consider the test error: $\mathcal{L}(A|S, f)$ (A = algorithm, S = training sample, f = ground truth) = $\sum_h \sum_{x\in\mathcal{X}\setminus S} P(x)I(h(x) \neq f(x))P(h|S, A)$. For binary classification, $f, h \in \mathcal{F} = \{f : \mathcal{X} \to \{0, 1\}\}$ where $|\mathcal{F}| = 2^{|\mathcal{X}|}$. Assume all fs are equally possible, so $\sum_f \mathcal{L}(A|S, f) = \sum_f \sum_h \sum_{x\in\mathcal{X}\setminus S} P(x)I(h(x) \neq f(x))P(h|S, A) = \sum_{x\in\mathcal{X}\setminus S} P(x) \sum_h P(h|S, A) \sum_f I(h(x) \neq f(x)) = 2^{|\mathcal{X}|-1} \sum_{x\in\mathcal{X}\setminus S} P(x)$. Here, we find alogorithm A is no longer related.

2 PAC/Nonasymptotic Framework

- In statistics, asymptotic includes consistency ($\hat{\theta}_n \to \theta_0$ *a.s.* as $n \to \infty$) and asymptotic normality (distribution) ($\sqrt{n}(\hat{\theta}_n - \theta_0) \to N(0, \Sigma)$) and its probability tools include LLNs and CLTs. Nonasymptotic includes error bounds with high probability and its probability tools include concentrating inequalities, large dimension bounds and empirical processes. In ML, PAC refers to Probably Approximately Correct = nonasymptotic + algorithm complexity.
  Chernoff bounds: $X_1, \cdots, X_n \in \{0, 1\}, \mu = E(\sum_{i=1}^n X_i)$. Then $\forall \alpha > 0$,
  1) $P(\sum_{i=1}^n X_i \geq (1+\alpha)\mu) < e^{-\frac{\mu\alpha^2}{2}}$; 2) $P(\sum_{i=1}^n X_i \leq (1-\alpha)\mu) < e^{-\frac{\mu\alpha^2}{2}}$.

- PAC framework
  Concept $c \in C : \mathcal{X} \to \mathcal{Y}$ is true parameter, and hypothesis $h \in \mathcal{H}$ is your estimation. Sample $S = (x_1, \cdots, x_m)$ with labels $(c(x_1), \cdots, c(x_m))$. Obtain h_S.
  Generalization error/risk $\mathcal{R}(h) = P_{x \sim D}(h(x) \neq c(x)) = E_{x \sim D} I(h(x) \neq c(x))$.
  Empirical error $\hat{\mathcal{R}}_S(h) = \frac{1}{m}\sum_{i=1}^m I(h(x_i) \neq c(x_i))$. Note that $\hat{\mathcal{R}}_S(h)$ is an unbiased estimater of $\mathcal{R}(h)$.
  Def: A concept class C is PAC-learnable if $\exists$ algorithm A & polynomial function $(\cdot, \cdot, \cdot, \cdot)$ s.t. $\forall \epsilon > 0, \delta > 0$,distribution D on $\mathcal{X}$,target concept c, it holds that $\forall m \geq \text{poly}(1/\epsilon, 1/\delta, n, \text{size}(c)), P_{S \sim D^m}$ $(\mathcal{R}(h_S) \leq \epsilon) \geq 1 - \delta$. If A runs in $\text{poly}(1/\epsilon, 1/\delta, n, \text{size}(c))$ then efficiently PAC-learnable.
  Computational costs: n: representation of $x \in \mathcal{X}$. size(c): representation of $c \in C$. They are usually dependent on dim($\mathcal{X}$). That is, sample + time + space.
  Example: $\mathcal{X} \in R^2, C = \{[l, r] \times [b, t] \,|\, l, r, b, t \in R\}$, true concept $R \in C, R_S \in C$ returned by the algorithm: tighted rectangle combining all points labeled 1.
  Fix $\epsilon > 0$, if $\epsilon \geq P(R)$, trivially, $\mathcal{R}(R_S) \leq P(R) \leq \epsilon$. Otherwise, assume $\epsilon < P(R)$. Find rectangles r_1, r_2, r_3, r_4 along the sides of R and $P(r_i) \geq \epsilon/4, P(\bar{r}_i(\text{excluding inner side})) \leq \epsilon/4$. If R_S intersects all r_is, then $\mathcal{R}(R_S) = P(R \backslash R_S) \leq 4 \times (\epsilon/4) = \epsilon$. Thus, $P_{S \sim D^m}(\mathcal{R}(R_S) > \epsilon) \leq \sum_{i=1}^4 P(R_S \cap r_i = \emptyset) \leq 4(1 - \epsilon/4)^m \leq 4e^{-m\epsilon/4}$. For m yields $m \geq \frac{4}{\epsilon}\log\frac{4}{\delta}$. Computation costs $O(1)$, time comlexity $O(m)$.
  Remark: Generlization bound: with probability $\geq 1 - \delta, \mathcal{R}(R_S) \leq \frac{4}{m}\log\frac{4}{\delta}$, i.e. $O(\frac{1}{m})$. Why the sharp rate? $\mathcal{H} = C$ is too simple!

- Finite hypothesis sets
  Consistent case: Assume A returns h_S s.t. $\hat{\mathcal{R}}_{\mathcal{S}}(h_S) = 0$.
  Thm: $\forall \epsilon, \delta > 0$, the inequality $P_{S \sim D}(\mathcal{R}(h_S) \leq \epsilon) \geq 1 - \delta$ holds if $m \geq \frac{1}{\epsilon}(\log|\mathcal{H}| + \log(\frac{1}{\delta}))$ or with probability $\geq 1 - \delta, \mathcal{R}(h_S) \leq \frac{1}{m}(\log(|\mathcal{H}| + \log(\frac{1}{\delta}))$.
  Proof: $\forall \epsilon > 0$, define $H_\epsilon = \{h \in \mathcal{H} : \mathcal{R}(h) > \epsilon\}$. Then $\forall h \in \mathcal{H}_\epsilon, P(\hat{\mathcal{R}}_S(h) = 0) \leq (1 - \epsilon)^m$. By the union bound, $P(\exists h \in \mathcal{H}_\epsilon \text{ s.t. } \hat{\mathcal{R}}_S(h) = 0) \leq \sum_{h \in \mathcal{H}_\epsilon} P(\hat{\mathcal{R}}_S(h) = 0) \leq |\mathcal{H}_\epsilon|(1 - \epsilon)^m \leq |\mathcal{H}|e^{-m\epsilon}$. Set RHS $= \delta$ and solve for m.
  Example: $C_n = (x_{j_1} \wedge \cdots \wedge x_{j_k} : k \leq n)$. e.g. $n = 4, c(x) = x_1 \wedge \bar{x}_2 \wedge x_4$. Consider the algorithm: for $(b_1, \cdots, b_n)$ labeled $+$, rule out $\bar{x}_i$ if $b_i = 1$, rule out x_i if $b_i = 0$. Also, $|\mathcal{H}| = |C_n| = 3^n$. Sample complexity $m \geq \frac{1}{\epsilon}(n\log 3 + \log(\frac{1}{\delta}))$. It is $O(n)$, which is greatly reduced from 3^n.
  Inconsistent case: $\forall h \in \mathcal{H}, \hat{\mathcal{R}}_S(h_S) \neq 0$.

Hoeffding inequality: $X_1,\cdots,X_n,\ X_i\in[a_i,b_i]$. Let $S_m=\sum_{i=1}^m x_i$, then $P(S_m-ES_m\ge\epsilon)\le\exp(-\frac{2\epsilon^2}{\sum_{i=1}^m(b_i-a_i)^2})$.
Corollary: $\forall\epsilon>0\,\&\,h:\mathcal{X}\to\{0,1\}\,,P_{S\sim D^m}(|\hat{\mathcal{R}}_S(h)-\mathcal{R}(h)|\ge\epsilon)\le 2e^{-2m\epsilon^2}$.
Setting RHS $=\delta$ and solving for ϵ yields for a single fixed $h:\forall\delta>0$, with probability $\ge 1-\delta$ it holds that $\mathcal{R}(h)\le\hat{\mathcal{R}}_S(h)+\sqrt{\frac{1}{2m}\log(\frac{2}{\delta})}$.
Thm: For finite $\mathcal{H},\forall$ choose a $h\in\mathcal{H},\cdots,\mathcal{R}(h)\le\hat{\mathcal{R}}_S(h)+\sqrt{\frac{1}{2m}(\log|\mathcal{H}|+\log(\frac{2}{\delta}))}$.
Remark: Choose the optimal $|\mathcal{H}|$, tradeoff between $\hat{\mathcal{R}}_S(h)\downarrow$ and $\log|\mathcal{H}|\uparrow$ as $|\mathcal{H}|\uparrow$.
Consider stochastic y: $S=((x_1,y_1),\cdots,(x_m,y_m))\sim D^m, D$ distribution on $\mathcal{X}\times\mathcal{Y}$. Generalization error $\mathcal{R}(h)=P_{(x,y)\sim D}(h(x)\neq y)=E_{(x,y)\sim D}I(h(x)\neq y)$.
Agnostic PAC learning: $\cdots,P_{S\sim D^m}(\mathcal{R}(h_S)-\min_{h\in\mathcal{H}}\mathcal{R}(h)\le\epsilon)\ge 1-\delta$.
Bayes classifier: h with the Bayes error $\mathcal{R}^*=\inf_{\text{measurable }h}\mathcal{R}(h)$. Why? $\forall x\in\mathcal{X},h_{\text{Bayes}}(x)=\text{argmax}_{y=0,1}P(y|x)(\text{unknown})$ with $\mathcal{R}^*=E\min(P(0|x),P(1|x))(\text{noise}(x))$. E[noise$(x)$] $=\mathcal{R}^*$.

- Infinte hypothesis sets (Rademacher complexity)
  Let g be a family of loss functions from $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$ to R, $G=\{g:z=(x,y)\to\mathcal{L}(h(x),y),h\in\mathcal{H}\}$.
  Empirical Rademacher complexity: $\hat{\text{Rad}}_S(G)=E_\sigma\sup_{g\in G}\frac{1}{m}\sum_{i=1}^m\sigma_i g(z_i)$ where σ_i i.i.d with $P(\sigma_i=\pm1)=\frac{1}{2}$.
  Rademacher complexity: $\text{Rad}_m(G)=E_{S\sim D^m}\hat{\text{Rad}}_S(G)$.
  Rationale: $\hat{\text{Rad}}_S(G)=E_\sigma\sup_{g\in G}\frac{<\sigma,g_S>}{m}$.
  McDiarmid inequality: If $\exists c_1,\cdots,c_m>0$ s.t. $|f(x_1,\cdots,x_i,\cdots,x_m)-f(x_1,\cdots,x_i',\cdots,x_m)|\le c_i\,\forall i$, then $\forall\epsilon>0,P(f(S)-Ef(S)\ge\epsilon)\le\exp(-\frac{2\epsilon^2}{\sum_{i=1}^m c_i^2})$.
  Thm: G is family of functions from $\mathcal{Z}$ to [0,1]. $\forall\delta>0$, with probability $\ge1-\delta$, it holds that $\forall g\in G:Eg(z)\le\frac{1}{m}\sum_{i=1}^m g(z_i)+2\text{Rad}_m(G)+\sqrt{\frac{\log(1/\delta)}{2m}},Eg(z)\le\frac{1}{m}\sum_{i=1}^m g(z_i)+2\text{Rad}_S(G)+3\sqrt{\frac{\log(2/\delta)}{2m}}$.
  Proof: Let $\Phi(S)=\sup_{g\in G}(Eg-\hat{E}_Sg)$. $\forall S,S'$ differing by exactly one point (z_i,z_i'). $\Phi(S)-\Phi(S')\le\sup_{g\in G}(\hat{E}_{S'}g-\hat{E}_Sg)=\sup_{g\in G}\frac{g(z_i')-g(z_i)}{m}\le\frac{1}{m}$. By McDiarmid inequality, with probability $\ge1-\delta,\Phi(S)\le E_S\Phi(S)+\sqrt{\frac{\log(1/\delta)}{2m}}$. We next bound $E_S\Phi(S)=E_S\sup_{g\in G}E_{S'}(\hat{E}_{S'}g-\hat{E}_Sg)\le E_{S,S'}\sup_{g\in G}(\hat{E}_{S'}g-\hat{E}_Sg)=E_{S,S'}\sup_{g\in G}\frac{1}{m}\sum_{i=1}^m(g(z_i')-g(z_i))=E_{\sigma,S,S'}\sup_{g\in G}\frac{1}{m}\sum_{i=1}^m\sigma_i(g(z_i')-g(z_i))\le E_{\sigma,S'}\sup_{g\in G}\frac{1}{m}\sum_{i=1}^m\sigma_ig(z_i')+E_{\sigma,S}\sup_{g\in G}\frac{1}{m}\sum_{i=1}^m-\sigma_ig(z_i)=2\text{Rad}_m(G)$. By McDiarmid inequality, with probability $\ge1-\delta/2,\text{Rad}_m(G)\le\hat{\text{Rad}}_S(G)+\sqrt{\frac{\log(2/\delta)}{2m}}$. Concluded by the union bound, with probability $\ge1-\delta/2,Eg\le\hat{E}_Sg+2\text{Rad}_m(G)+\sqrt{\frac{\log(2/\delta)}{2m}}$.
  Lemma: $h\in H$ taking values in $\{-1,1\}\,,G=\{(x,y)\to I(h(x)\neq y):h\in H\}$, then $\hat{\text{Rad}}_S(G)=\frac{1}{2}\text{Rad}_{S_\mathcal{X}}(H)$.
  Proof: $\hat{\text{Rad}}_S(G)=E_\sigma\sup_{h\in H}\frac{1}{m}\sum_{i=1}^m\sigma_iI(h(x_i)\neq y_i)=E_\sigma\sup_{h\in H}\frac{1}{m}\sum_{i=1}^m\sigma_i\frac{1-y_ih(x_i)}{2}=\frac{1}{2}E_\sigma\sup_{h\in H}\frac{1}{m}\sum_{i=1}^m\sigma_ih(x_i)=\frac{1}{2}\hat{\text{Rad}}_{S_\mathcal{X}}(H)$.
  Thm: H is family of functions taking values in $\{-1,1\}$. $\forall\delta>0$, with probability $\ge1-\delta$, it holds that $\forall h\in H:\mathcal{R}(h)\le\hat{\mathcal{R}}_S(h)+\text{Rad}_m(H)+\sqrt{\frac{\log(1/\delta)}{2m}},\mathcal{R}(h)\le\hat{\mathcal{R}}_S(h)+\text{Rad}_S(H)+3\sqrt{\frac{\log(2/\delta)}{2m}}$.

- Infinte hypothesis sets (VC dimension)
  Growth function: $\Pi_\mathcal{H}(m)=\max_{\{x_1,\cdots,x_m\}\subset\mathcal{X}}|\{(h(x_1),\cdots,h(x_m)):h\in\mathcal{H}\}|$.
  Lemma (Massart): $A\subset R^m$ a finite set, $r=\max_{x\in A}||x||_2$, then $E_\sigma(\frac{1}{m}\sup_{x\in A}\sum_{i=1}^m\sigma_ix_i)\le\frac{r\sqrt{2\log|A|}}{m}$ where $x=(x_1,\cdots,x_m)$.

Proof: This follows from the maximal inequality: If $X_1, \cdots, X_n$ are sub-Gaussian, i.e. $\exists r > 0$ s.t. $Ee^{tX_j} \leq e^{t^2r^2/2}, \forall t > 0$, then $E\text{max}_{1\leq j\leq n}X_j \leq r\sqrt{2\text{log}n}$. To show this, $\forall t > 0$, by Jensen's inequality, $e^{tE\text{max}_j X_j} \leq Ee^{t\text{max}_j X_j} \leq \sum_{i=1}^n Ee^{tX_j} \leq ne^{t^2r^2}/2$. Taking log and dividing by t, we have $E\text{max}_j X_j \leq \frac{\text{log}n}{t} + \frac{tr^2}{2}$. Choosing $t = \sqrt{2\text{log}n}/r$ yields the inequality. Check $\sum_{i=1}^m \sigma_i x_i \sim$ sub-G with parameter $r = ||x||_2$.

Corollary: $g \in G$ taking values in $\{-1, 1\}$, then $\text{Rad}_m(G) \leq \sqrt{\frac{2\text{log}\Pi_G(m)}{m}}$.

Proof: $\text{Rad}_m(G) = E_S E_\sigma \text{sup}_{g\in G}\frac{1}{m}\sum_{i=1}^m \sigma_i g(x_i) \leq \frac{\sqrt{m}\sqrt{2\text{log}|(g(x_1),\cdots,g(x_m)):g\in G|}}{m} \leq \sqrt{\frac{2\text{log}\Pi_G(m)}{m}}$.

VC dimension: The VC dimension of $\mathcal{H}$ in the size of the largest set that can be shattered by $\mathcal{H}$: VCdim($\mathcal{H}$)=max$\{m : \Pi_{\mathcal{H}}(m) = 2^m\}$.

Computing VCdim: 1) $\exists$ a set of size d shattered by $\mathcal{H}$; 2) no set of size $d+1$ can be shattered by $\mathcal{H}$.

In R^d, VCdim $= d+1$, $S = (0, e_1, \cdots, e_d)$ with arbitrary labels $y = (y_0, y_1, \cdots, y_d)$. Take hyperplain $(y, x) + \frac{y_0}{2} = 0$. sgn($(y, 0) + \frac{y_0}{2}$)=$y_0$, sgn($(y, e_i) + \frac{y_0}{2}$)=$y_i, \forall i = 1, \cdots, d$.

Radon's thm: Any set of $d+2$ points in R^d can be partioned into two subsets where convex hulls intersect.

Bounding growth function via VCdim: If VCdim(H) $= \infty$, $\Pi_{\mathcal{H}}(m) = 2^m$. What if VCdim($\mathcal{H}$) $= d < \infty$?

Sauel's lemma: If VCdim($\mathcal{H}$) $= d$, then $\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^d C_m^i$. Proof: By induction on $m + d$.

Corollary: $\Pi_{\mathcal{H}}(m) \leq (\frac{em}{d})^d = O(m^d), \forall m \geq d$.

Proof: By Sauel's lemma, if $m \geq d, \Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^d C_m^i \leq \sum_{i=0}^d C_m^i(\frac{m}{d})^{d-i} \leq \sum_{i=0}^m C_m^i(\frac{m}{d})^{d-i} = (\frac{m}{d})^d\sum_{i=0}^m C_m^i(\frac{d}{m})^i = (\frac{m}{d})^d(1+\frac{d}{m})^m \leq (\frac{m}{d})^d e^d$.

Corollary: $h \in \mathcal{H}$ taking values in $\{-1, 1\}$, VCdim($\mathcal{H}$) $= d$, then we have generalization bound: $\mathcal{R}(h) \leq \hat{\mathcal{R}}_S(h) + \sqrt{\frac{2d\text{log}(em/d)}{m}} + \sqrt{\frac{\text{log}(1/\delta)}{2m}}$ with probability at least $1 - \delta$.

- Lower bounds

  Realizable: $\exists h \in \mathcal{H}$ s.t. $\mathcal{R}(h) = 0$.

  Idea: Probabilistic method: Find a bad distribution for any algorithm.

  Thm: If VCdim($\mathcal{H}$) $= d > 1$, then $\forall m \geq 1$ and algorithm, $\exists$ distribution D and target concept $f \in \mathcal{H}$, s.t. $P_{S\sim D^m}(\mathcal{R}_D(h_S, f) > \frac{d-1}{32m}) \geq 0.01$.

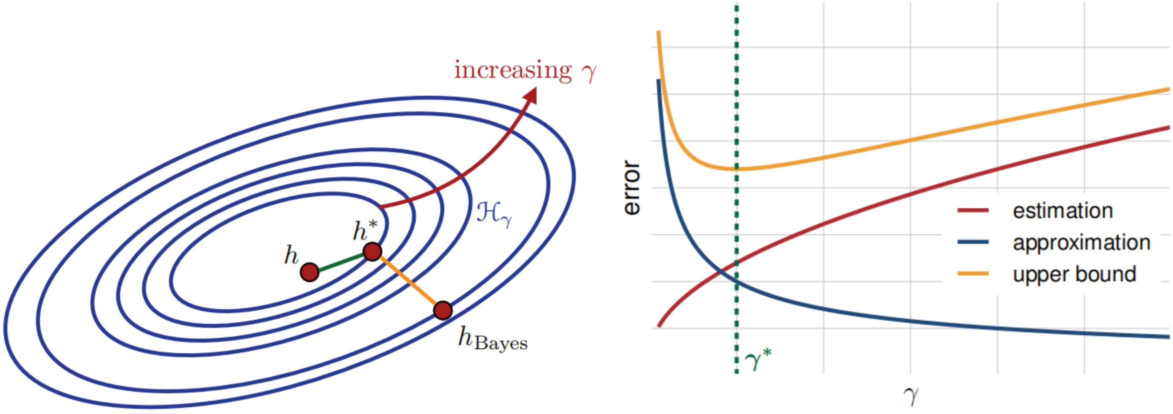
  Proof: Assume $\bar{X} = (x_0, x_1, \cdots, x_{d-1}) \subset \mathcal{X}$ be shattered by $\mathcal{H}$. $\forall\epsilon > 0$, choose D s.t. $P_D(x_0) = 1 - 8\epsilon, P_D(x_i) = \frac{8\epsilon}{d-1}$. WLOG, let's assume A makes no error on x_0. $\forall S$, we let $\bar{S}$ be the set of S's elements in $\{x_1, \cdots, x_{d-1}\}$ and $\mathcal{S}$ be the set of S of size m satisfying $|\bar{S}| \leq \frac{d-1}{2}$. Then fix $S \in \mathcal{S}$ and let U be the uniform distribution over all labelings $f : \bar{X} \to \{0, 1\}$. Thus $E_{f\sim U}\mathcal{R}_D(h_S, f) = \sum_f \sum_{x\in\bar{X}} I(h_S(x) \neq f(x))P(x)P(f) \geq \sum_{x\in\bar{X}\setminus\bar{S}} P(x) \sum_f I(h_S(x) \neq f(x))P(f) \geq \frac{1}{2}\sum_{x\notin\bar{S}} P(x) \geq \frac{1}{2}\frac{d-1}{2}\frac{8\epsilon}{d-1} = 2\epsilon$. Taking expectation over $S \in \mathcal{S}$ by Fubini's theorem, $E_{f\sim U}E_{S\in\mathcal{S}}\mathcal{R}_D(h_S, f) \geq 2\epsilon$. Thus, $\exists f_0 \in \mathcal{H}$ s.t. $E_{S\sim\mathcal{S}}\mathcal{R}_D(h_S, f_0) \geq 2\epsilon$. We can notice that $\mathcal{R}_D(h_S, f_0) \leq P_D(\bar{X} - \{x_0\}) \leq 8\epsilon$, thus $2\epsilon \leq E_{S\sim\mathcal{S}}\mathcal{R}_D(h_S, f_0) \leq 8\epsilon P_{S\in\mathcal{S}}(\mathcal{R}_D(h_S, f_0) \geq \epsilon) + \epsilon P_{S\in\mathcal{S}}(\mathcal{R}_D(h_S, f_0) < \epsilon)$. Solving yields $P_{S\in\mathcal{S}}(\mathcal{R}_D(h_S, f_0) \geq \epsilon) \geq \frac{1}{7} \Rightarrow P_S(\cdots) \geq \frac{1}{7}P(\mathcal{S})$. Finally, by the multiplicative Chernoff bound, $P(|\bar{S}| \geq 8\epsilon m(1+\gamma)) \leq e^{-8\epsilon m\gamma^2/3}, \forall\gamma > 0$. Taking $\epsilon = \frac{d-1}{32m}$ and $\gamma = 1$, $P(|\bar{S}| \geq \frac{d-1}{2}) \leq e^{-(d-1)/12} \leq e^{-1/12}$. We conclude that $P_S(\mathcal{R}_D(h_S, f_0) \geq \epsilon) \geq \frac{1}{7}(1 - e^{-1/12}) > 0.01$.

3 Model Selection and Regularization

- Tradeoff – excess error

  $\mathcal{R}(h') - R^*$(Bayes optimal) $= [\mathcal{R}(h') - \inf_{h\in\mathcal{H}}\mathcal{R}(h)] + [\inf_{h\in\mathcal{H}}\mathcal{R}(h) - \mathcal{R}^*]$. The former is estimation(statistical) error, and the latter is approximation error. Usuallt, take a rich family $\mathcal{H} = \cup_{\gamma\in\Gamma}\mathcal{H}_\gamma$, as $\gamma\uparrow$, estimation error $\uparrow$, approximation error $\downarrow$.

  In particular, $y = f(x) + \epsilon, E(\epsilon|X) = 0, \mathrm{Var}(\epsilon|X) = \sigma^2$. Under squared error loss, $\mathcal{R}(x_0) = E[(y - \hat{f}(x_0))^2|X = x_0] = \sigma^2 + [E\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - E\hat{f}(x_0)]^2$. 1st is irreducible error, 2nd is bias^2, 3rd is variance. Similarly, as $\gamma\uparrow$, bias $\downarrow$, variance $\uparrow$.



  For the LHS, the estimation error is in green and approximation error is in orange. For the RHS, leftside of γ^* is underfit and rightside is overfit.

- General Methodology

  Empirical risk minimization (ERM): $h_S^{\text{ERM}} = \text{argmin}_{h\in\mathcal{H}}\hat{R}_S(h)$.

  Estimation error of ERM: $P(\mathcal{R}(h_S^{\text{ERM}}) - \inf_{h\in\mathcal{H}}\mathcal{R}(h) > \epsilon) \le P(\sup_{h\in\mathcal{H}}|\mathcal{R}(h) - \hat{\mathcal{R}}_S(h)| > \frac{\epsilon}{2})$.

  Proof: $\forall\epsilon > 0, \exists h_\epsilon$ s.t. $\mathcal{R}(h_\epsilon) \le \inf_{h\in\mathcal{H}}\mathcal{R}(h) + \epsilon$. Then $\mathcal{R}(h_S^{\text{ERM}}) - \inf_{h\in\mathcal{H}}\mathcal{R}(h) = \mathcal{R}(h_S^{\text{ERM}}) - \hat{\mathcal{R}}_S(h_S^{\text{ERM}}) + \hat{\mathcal{R}}_S(h_S^{\text{ERM}}) - \mathcal{R}(h_\epsilon) + \mathcal{R}(h_\epsilon) - \inf_{h\in\mathcal{H}}\mathcal{R}(h) \le 2\sup_{h\in\mathcal{H}}|\mathcal{R}(h) - \hat{\mathcal{R}}(h)| + \epsilon$. Letting $\epsilon\to 0$ yields $\mathcal{R}(h_S^{\text{ERM}}) - \inf_{h\in\mathcal{H}}\mathcal{R}(h) \le 2\sup_{h\in\mathcal{H}}|\mathcal{R}(h) - \hat{\mathcal{R}}(h)|$.

  Remark: Previous shown that with high probability $1-\delta$, $\sup_{h\in\mathcal{H}}|\mathcal{R}(h) - \hat{\mathcal{R}}(h)| \le \mathrm{Rad}_m(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2m}}$. ERM does not work for too rich families.

  Structure risk minimization (SRM): $\mathcal{H} = \cup_{k=1}^\infty\mathcal{H}_k$ s.t. $\mathcal{H}_k \subset \mathcal{H}_{k+1}, h_S^{\text{SRM}} = \text{argmin}_{k\ge 1, h\in\mathcal{H}_k}\hat{\mathcal{R}}_S(h) + \mathrm{Rad}_m(\mathcal{H}_k) + \sqrt{\frac{\text{log}k}{m}}$. The second term can be replaced by other complexity measures and the last term of inflation is due to multiple choices.

  Thm: with high probability, $\mathcal{R}(h_S^{\text{SRM}}) \le \inf_{h\in\mathcal{H}}[\mathcal{R}(h) + 2\mathrm{Rad}_m(\mathcal{H}_k) + \sqrt{\frac{\text{log}k(h)}{m}}] + \sqrt{\frac{2\text{log}(3/\delta)}{m}}$.

- Regularization

  Constrained: $\text{argmin}_{\gamma>0, h\in\mathcal{H}_\gamma}\hat{\mathcal{R}}(h) + \mathrm{Pen}(\gamma, m)$.

  Unconstrained: $\text{argmin}_{h\in\mathcal{H}_\gamma}\hat{\mathcal{R}}(h) + \lambda\mathrm{Pen}(h)$.

  Here Pen refers to Penalty Function. The result is $\lambda\to 0$: ERM; $\lambda\uparrow$: h simpler.

- Cross-Validation

  For choosing the final model. Denote traning set as S_1, validation set S_2. $|S_1| = (1-\alpha)m, |S_2| =$

$\alpha m, \alpha \in (0,1)$. $h_S^{\mathrm{CV}} = \mathrm{argmin}_{h \in \left\{h_{S_1}^{\mathrm{ERM}}, k \geq 1\right\}} \hat{\mathcal{R}}_{S_2}(h)$.

Prop: $P(\sup_{k \geq 1} |\mathcal{R}(h_{S_1,k}^{\mathrm{ERM}}) - \hat{\mathcal{R}}_{S_2}(h_{S_1,k}^{\mathrm{ERM}}) - \sqrt{\frac{\mathrm{log}k}{\alpha m}}| > \epsilon) \leq 4e^{-2\alpha m \epsilon^2}$.

Proof: By the union bound and law of total expectation, LHS $\leq \sum_{k=1}^{\infty} EP(|\mathcal{R}(h_{S_1,k}^{\mathrm{ERM}}) - \hat{\mathcal{R}}_{S_2}(h_{S_1,k}^{\mathrm{ERM}})| > \epsilon + \sqrt{\frac{\mathrm{log}k}{\alpha m}} | S_1) \leq \sum_{k=1}^{\infty} 2e^{-2\alpha m \epsilon^2 - 2\mathrm{log}k} = \sum_{k=1}^{\infty} \frac{2}{k^2} e^{-2\alpha m \epsilon^2} = \frac{\pi^2}{3} e^{-2\alpha m \epsilon^2} \leq 4e^{-2\alpha m \epsilon^2}$.

Thm: w.h.p., $\mathcal{R}(h_S^{\mathrm{CV}}) - \mathcal{R}(h_S^{\mathrm{SRM}}) \leq 2\sqrt{\frac{\log\ \max(k(h_S^{\mathrm{CV}}), k(h_{S_1}^{\mathrm{SRM}}))}{\alpha m}} + 2\sqrt{\frac{\log(4/\delta)}{2\alpha m}}$.

K-fold CV: $\hat{\mathcal{R}}_{\mathrm{CV}}(\theta) = \frac{1}{K} \sum_{i=1}^{K} \frac{1}{m_i} \sum_{j=1}^{m_i} \mathcal{L}(h_i(x_{ij}), y_{ij})$.

Tradeoff: as $k \uparrow$, we have $m_i \downarrow$, bias $\downarrow$, variance $\uparrow$.

$K = m \rightarrow$ leave-one-out CV. Usually, $k = 5$ or 10.

- Information Criterion

Generalization error = Training error + "Optimism"

Def $\hat{\mathcal{R}} = \hat{\mathcal{R}}(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{f}(x_i)), \mathcal{R}_T = \mathcal{R}_T(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} E_{y^0}(\mathcal{L}(y_i^0, \hat{f}(x_i))|T)$ (in-sample test error),Opt $= \mathcal{R}_\tau - \hat{\mathcal{R}}$. Here we consider $y_i^0 = f(x_i) + \epsilon_i^0, \mathrm{Var}(\epsilon) = \sigma^2 I_n$.

Under squared error loss, $E(y_i - \hat{y}_i)^2 = \mathrm{Var}(y_i) + \mathrm{Var}(\hat{y}_i) - 2\mathrm{Cov}(y_i, \hat{y}_i) + (Ey_i - E\hat{y}_i)^2, E(y_i^0 - \hat{y}_i)^2 = \mathrm{Var}(y_i^0) + \mathrm{Var}(\hat{y}_i) - 2\mathrm{Cov}(y_i^0, \hat{y}_i) + (Ey_i^0 - E\hat{y}_i)^2$. Thus, $E(\mathrm{opt}) = \frac{2}{n} \sum_{i=1}^{n} \mathrm{Cov}(y_i, \hat{y}_i)$.

Consider linear fit $\hat{y} = Hy$, thus $\mathrm{Cov}(y_i, \hat{y}_i) = \mathrm{Cov}(e_i^T y, e_i^T Hy) = e_i^T \mathrm{Var}(y) He_i = \sigma^2 e_i^T He_i = \sigma^2 h_{ii}$, so that $E(\mathrm{opt}) = \frac{2}{n} \sigma^2 \mathrm{tr}(H)$.

Estimate in-sample test error: Mallow's C_P: $C_p = \hat{\mathcal{R}} + 2\frac{d}{n}\hat{\sigma}^2$ (d refers to degrees of freedom).

Generalize to log-likelihood: Akaike information criterion (AIC): AIC $= -2\mathrm{log}(\mathrm{lik}) + 2d$.

Bayesian information criterion (BIC) for model selection:

Given candidate models $M_1, \cdots, M_K$, model parameters $\theta_1, \cdots, \theta_K$, prior $\pi(M_j)$ on models and $p(\theta_j|M_j)$ on parameters. Posterior $p(M_j|T)$. The odds ratio: $\frac{p(M_j|T)}{p(M_k|T)} = \frac{\pi(M_j)}{\pi(M_k)} \frac{p(T|M_j)}{p(T|M_k)}$.

For model M, $p(T|M) = \int p(T|\theta) p(\theta|M) d\theta$, so that $-\mathrm{log}p(T|M) = -\mathrm{log} \int p(T|\theta) p(\theta|M) d\theta := -\mathrm{log} \int e^{L_M(\theta)} d\theta$. By Taylor expansion, $L_M(\theta) \approx L_M(\hat{\theta}) + (\theta - \hat{\theta})^T \nabla L_M(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^T \nabla^2 L_M(\hat{\theta})(\theta - \hat{\theta})$. Taking $\hat{\theta}$ as the MAP estimation s.t.$\nabla L_M(\hat{\theta}) = 0$. By substituting back, $\int e^{L_M(\theta)} d\theta \approx \int p(T|\hat{\theta}) p(\hat{\theta}|M) \mathrm{exp}(-\frac{1}{2}(\theta - \hat{\theta})^T n\hat{I}(\hat{\theta})(\theta - \hat{\theta})) d\theta = p(T|\hat{\theta}) p(\hat{\theta}|M) (2\pi)^{\frac{d}{2}} \det(n\hat{I}(\hat{\theta}))^{-\frac{1}{2}}$. Thus, $-\mathrm{log}p(T|M) = -\mathrm{log}p(T|\hat{\theta}) - \mathrm{log}p(\hat{\theta}|M) - \frac{d}{2}\mathrm{log}(2\pi) + \frac{d}{2}\mathrm{log}n + \frac{1}{2}\mathrm{logdet}(\hat{I}(\hat{\theta}))$ or $-2\mathrm{log}p(T|M) = -2\mathrm{log}(\mathrm{lik}) + (\mathrm{log}n)d$ (namely BIC).

AIC vs BIC: 1) AIC denser, BIC sparser; 2) Theoretical guarantees: AIC minimax rate-optimal for prediction, BIC consistent for model selection; 3) AIC-BIC dilemma.

Bootstrap method for direct estimation of generalization error:

Goal: "out-of-sample" test error $\mathcal{R} = E\mathcal{L}(y, \hat{f}(x))$. Training sample $Z = \{(x_i, y_i) : i = 1, \cdots, n\}$.

Generate B bootstrap samples with replacement: $Z^{*(1)}, \cdots, Z^{*(B)}$.

Native estimate: $\hat{\mathcal{R}}_{\mathrm{boot}} = \frac{1}{Bn} \sum_{b=1}^{B} \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{f}^{*(b)}(x_i))$ does not work.

Exclude bootstrap samples containing obs: $\mathcal{R}_{\mathrm{boot}}^{(1)} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|C_{-i}|} \sum_{b \in C_{-i}} \mathcal{L}(y_i, \hat{f}^{*(b)}(x_i))$.

Q: How many distinct obs are used in training?

P(obs i sampled) $= 1 - (1 - \frac{1}{n})^n \rightarrow 1 - e^{-1} \approx 0.632$.

Upward bias due to smaller training sample size.

Remedy: 1) .632 estimate: $\hat{\mathcal{R}}^{(.632)} = .368\hat{\mathcal{R}} + .632\hat{\mathcal{R}}^{(1)}$;

2) .632+ estimate: $\hat{\mathcal{R}}^{(.632+)} = (1 - \hat{\omega})\hat{\mathcal{R}} + \hat{\omega}\hat{\mathcal{R}}^{(1)}$ where $\hat{\omega}$ is data-driven weight.

4 Linear Regression and Classification

- Linear Regression

  Given the covariate/predictor $x \in R^p$ and response $y \in R$, $E(Y|X) = X^T\beta = \beta_0 + \sum_{j=1}^{p-1} x_j\beta_j$ where $x_0 = 1, x = (x_0, \cdots, x_{p-1})^T, \beta = (\beta_0, \cdots, \beta_{p-1})$. Data $\{(x_i, y_i)\}_{i=1}^n$ i.i.d. The linear model: $Y = X\beta + \epsilon$.

  Least squares: minimise $\text{RSS}(\beta) = ||Y - X\beta||_2^2$. Differentiating, $X^T(Y - X\beta) = 0 \Leftrightarrow X^TX\beta = X^TY$. If X has full column rank, then $\hat\beta = (X^TX)^{-1}X^TY$. Predicted values: $\hat f(x_0) = x_0^T\hat\beta$. Otherwise, less than full rank: $\hat y$ is stil unique, but $\hat\beta$ is not.

  Remedy: 1) drop redundant variables; 2) regularization, especially when $p > n$.

  Sampling Properties: Assume fixed deterministic design X and $\text{Var}(\epsilon) = \sigma^2 I_n$. Then $E\hat\beta = \beta$ (unbiasedness), $\text{Var}(\hat\beta) = \text{Var}((X^TX)^{-1}X^TY) = (X^TX)^{-1}X^T\text{Var}(Y)X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$. If $\epsilon \sim N(0, \sigma^2)$, then $\hat\beta \sim N_p(\beta_0, \sigma^2(X^TX)^{-1}), \hat\sigma^2 = \frac{1}{n-p}||y - \hat y||_2^2. E\hat\sigma^2 = \sigma^2, (n-p)\frac{\hat\sigma^2}{\sigma^2} \sim \chi^2(n-p)$.

  Algorithms for computing LSE: Orthogonal design $(x_j, x_k) = 0, \forall j, k$. $\hat\beta_j = \frac{(y, x_j)}{(x_j, x_j)}, j = 1, \cdots, p$.

  Gram-Schmidt: 1) Initialize $z_0 = x_0 = \mathbf{1}$; 2) For $j = 1, \cdots, p$, regress x_j on $z_0, \cdots, z_{j-1}$ to get $\hat\gamma_{lj} = \frac{(z_l, x_j)}{(z_l, z_l)}, l = 0, \cdots, j-1$ and residual vector $z_j = x_j - \sum_{k=0}^{j-1}\hat\gamma_{kj}z_k$; 3) Regress y on the residual $z_p : \hat\beta_p = \frac{(y, z_p)}{(z_p, z_p)}$.

  Interpretation of $\hat\beta_j$: additional contribution of x_j to y after "adjusting" $x_0, \cdots, x_{j-1}, x_{j+1}, \cdots, x_p$.

  Effect of multicollinearity: $\text{Var}(\hat\beta_p) = \frac{\sigma^2}{||z_p||_2^2}$ large when $||z_p||_2$ is small.

  Remedy: regularization (later).

  Q: How to get all β_j in one pass?

  QR decomposition: $X = Z\Gamma = (z_0, \cdots, z_p)\begin{pmatrix} 1 & \hat\gamma_{01} & \cdots & \hat\gamma_{0p} \\ 0 & 1 & \cdots & \hat\gamma_{1p} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$. Let $D = \text{diag}(||z_0||_2, \cdots, ||z_p||_2)$, $X = (ZD^{-1})(D\hat\Gamma) = QR$ where $Q^TQ = I$. So $\hat\beta = (X^TX)^{-1}X^TY = R^{-1}Q^TY, \hat y = x\hat\beta = QQ^TY$.

  Mult-response/multi-task learning: $Y = XB + E$, $Y \in M_{n\times q}(R), X \in M_{n\times p}(R), B \in M_{p\times q}(R), E \in M_{n\times q}(R)$. Minimise $\text{RSS}(B) = ||Y - XB||_F^2 = \text{tr}((Y - XB)^T(Y - XB)), \hat B = (X^TX)^{-1}X^TY$, deconples to q unrelated regressions.

  Assume $e_i \sim N_g(0, \Sigma)$. MLE is equivalent to minimising $-2\text{log}(\text{lik}(B; \Sigma)) = \text{tr}((Y - XB)^T(Y - XB)\Sigma^{-1}) + \text{const}(\Sigma), \hat B(\Sigma) =$ the same.

  Q: When is the problem not separable?

  1) each eequation has different sets of predictors and and Σ is not diagonal;

  2) B has low rank (or other constraint/regularization).

- Linear Classification

  Decision boundaries on which $\hat f_k(x) = \hat f_l(x)$, i.e.$\hat f_k(x) = \hat\beta_{k0} + \hat\beta_k^T x$, $\left\{x|(\hat\beta_{k0} - \hat\beta_{l0}) + (\hat\beta_k - \hat\beta_l)^T x = 0\right\}$.

  Method 1: Linear discriminant analysis (LDA).

  $P(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^K f_l(x)\pi_l}$ where $f_l(x)$ is density for class l and π_l is prior of class l.

  Idea: Use Gaussian density with common Σ, $f_k(x) = (2\pi)^{-p/2}\text{det}(\Sigma)^{-1/2}\text{exp}(-\frac{1}{2}(x - \mu_k)^T\Sigma^{-1}(x - \mu_k))$. log-odds ratio: $\text{log}\frac{P(k|x)}{P(l|x)} = \text{log}\frac{\pi_k}{\pi_l} - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \frac{1}{2}\mu_l^T\Sigma^{-1}\mu_l + x^T\Sigma^{-1}(\mu_k - \mu_l)$. Here, $\Sigma^{-1}(\mu_k - \mu_l)$

is discriminant direction.

Discriminant function: $\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k$. Decision boundaries: $\{x|\delta_k(x) = \delta_l(x)\}$.

Fitting: $\hat{\pi}_k = \frac{N_k}{N}, \hat{\mu}_k = \sum_{g_i=k} x_i/N_k, \hat{\Sigma} = \sum_{k=1}^{K}\sum_{g_i=k}(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T/(N-K)$.

Extension: Quadratic Discriminant Analysis (QDA): Use class-specific Σ: $\hat{\Sigma}_k$ = sample covariance of class k. (No longer a linear method!)

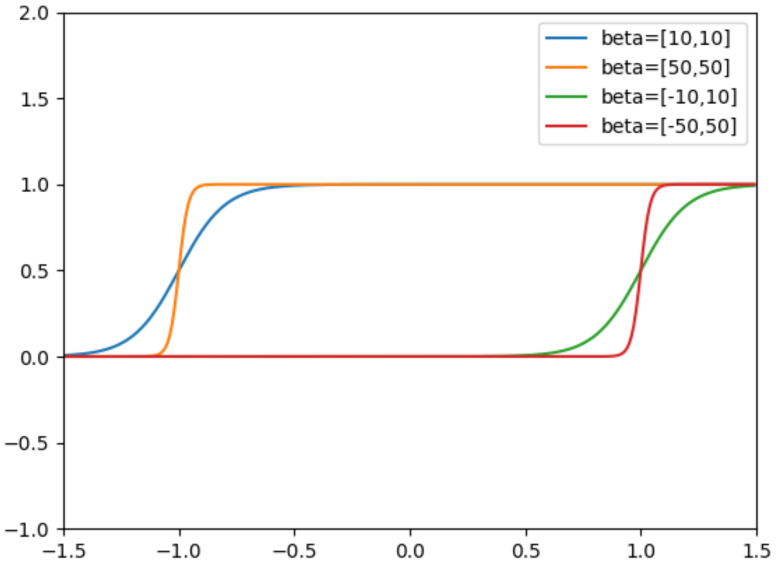
Method 2: Logistic regression.

Naive method: indicator response matrix $Y_{n\times k}$ with a single 1 in each row ("one hot"). Then $Y = XB + E, \hat{B} = (X^TX)^{-1}X^TY, \hat{f}(x_0) = \hat{B}^Tx_0, \hat{G}(x) = \text{argmax}_k\hat{f}_k(x)$.

Logistic: $\log\frac{P(k|x)}{P(K|x)} = \beta_{0k} + \beta_k^Tx, k = 1,\cdots,K-1$. Solving yields $P(k|x) = \frac{\exp(\beta_{0k}+\beta_k^Tx)}{1+\sum_{l=1}^{K-1}\exp(\beta_{0l}+\beta_l^Tx)}, k = 1,\cdots,K-1, P(K|x) = \frac{1}{1+\sum_{l=1}^{K-1}\exp(\beta_{0l}+\beta_l^Tx)}$.

Remark: Permutation invariant, but not for structured B.

When $K = 2$, logistic regression has a strong bond with sigmoid function. Specifically, it conducts a linear transform over x of Sigmoid(x) : $\frac{1}{1+e^{-x}} \to \frac{1}{1+e^{-(\beta_0+\beta_1x)}}$.



Fitting MLE via Newton-Raphsen (iteratively reweighted least squares).

Specially, if samples of different classes can be linearly separated, MLE is undefined (i.e. infinite).

Comparing LDA and LR: In LDA, $\log\frac{P(k|x)}{P(K|x)} = \log\frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k-\mu_l)^T\Sigma^{-1}(\mu_k-\mu_l) + x^T\Sigma^{-1}(\mu_k-\mu_l) = \alpha_{0k} + \alpha_k^Tx$ while in LR, $\log\frac{P(k|x)}{P(K|x)} = \beta_{0k} + \beta_k^Tx$. They take the same form. A great difference is in $P(X,G) = P(X)P(G|X)$, where $P(X) \to$ Gaussian for LDA while $P(X) \to$ unspecified for LR.

- Some Propositions

  1) Consider Y is indicator matrix, regress X on Y and we get $\hat{B}, \hat{Y} = X\hat{B}$. Show that using LDA on X is equivalent to LDA on $\hat{Y}$. (Hint: prove $x^T\Sigma^{-1}\mu_k = \hat{y}^T\hat{\Sigma}^{-1}\hat{\mu}_k$ and $\mu_k^T\Sigma^{-1}\mu_k = \hat{\mu}_k^T\hat{\Sigma}^{-1}\hat{\mu}_k$)

  2) If $\text{RSS}(B,\Sigma) := \sum_{i=1}^{N}(y_i - f(x_i))^T\Sigma^{-1}(y_i - f(x_i))$, we also have the same estimate, i.e.$\hat{B} = (X^TX)^{-1}X^TY$. (Hint: by letting $\frac{\partial\text{RSS}(B,\Sigma)}{\partial B} = 0$)

  3) If $X = (X_1, X_1, \cdots, X_1)$ has identical columns, then by ridge regression we will get identical coefficients $\beta = (\beta_1, \beta_1, \cdots, \beta_1)$.

5 Lasso and Related Methods

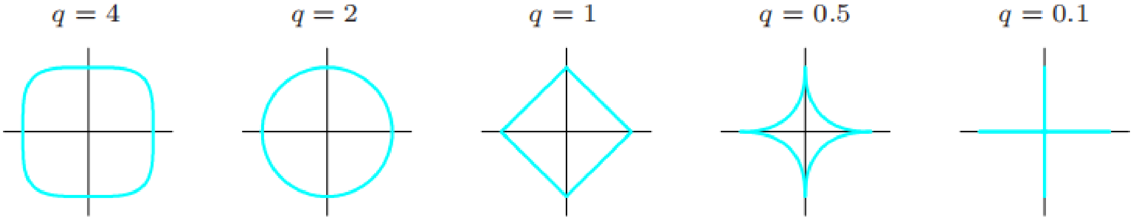
- Why imporve least squares estimate?

  1) generalzation: MSE = bias^2 + var, trade bias for variance;

  2) interpretation: sparsity, variable selection;

  3) stability: robust to perturbations on data.

  Contours of different $\mathscr{L}_q$ norms:

- Best subset selection

  Best subset selection ($\mathscr{L}_0$ regularization): minimise $||y - x\beta||_2^2$ s.t. $||\beta||_0 \le k$ where k is a tuning parameter and $||\beta||_0 := |\{j : \beta_j \neq 0\}|$.

  Remark: $\mathscr{L}_0$ regularization is hard to get accurate solutions.

  Algorithms: Mixed integer optimization. $p \sim$ 100s or 1000s for approximate solutions.

  Greedy strategies: stepwise selection: 1) forward: start with null model, add one prediction at a time; 2) backward: start with full model, delete one prediction at a time; 3) bidirectional: combine both forward and backward solution.

  Dropbacks: 1) Solution path not continuous, unstable; 2) No bias for $k \ge ||\beta^*||_0$, high variance.

- Ridge regression

  Ridge regression ($\mathscr{L}_2$ regularization, weight decay in deep learning): minimise $||y - x\beta||_2^2 + \lambda||\beta||_2^2 \Leftrightarrow$ minimise $||y - x\beta||_2^2$ s.t. $||\beta||_2 \le t$. ("$\Leftrightarrow$" is due to the convexity)

  Centering: we default there exists no intercept. Otherwise, center both y and x by $y - \bar{y}, x_{ij} - \bar{x}_j$. Solving yields $\hat{\beta} = (X^TX + \lambda I)^{-1}X^TY$.

  Shrinkage effect: SVD $X = U_{n\times p}D_{p\times p}V_{p\times p}^T$, U has orthogonal columns ($U^TU = I$), V orthogonal ($V^TV = VV^T = I$), $D = \text{diag}(d_1, \cdots, d_p)$. (Assume $n > p, d_1 \ge \cdots \ge d_p$)

  Least squares: $X\hat{\beta}^{\text{ls}} = X(X^TX)^{-1}X^TY = UU^TY = \sum_{j=1}^p u_ju_j^TY$.

  Ridge: $X\hat{\beta}^{\text{ridge}} = X(X^TX+\lambda I)^{-1}X^TY = UDV^T(VD^2V^T+\lambda I)^{-1}VDU^TY = UD^2(D^2+\lambda I)^{-1}U^TY = \sum_{j=1}^p \frac{d_j^2}{d_j^2+\lambda}u_ju_j^TY$. $\lambda > 0 \Rightarrow \frac{d_j^2}{d_j^2+\lambda} \le 1 \Rightarrow$ Shrinkage.

  Degree of freedom: $df(\lambda) = \text{tr}(X(X^TX + \lambda I)^{-1}X^T) = \text{tr}(UD^2(D^2 + \lambda I)^{-1}U^T) = \text{tr}(D^2(D^2 + \lambda I)^{-1}) = \sum_{i=1}^p \frac{d_j^2}{d_j^2+\lambda}$. $df(\lambda) \to p$ as $\lambda \to 0$ and $df(\lambda) \to 0$ as $\lambda \to \infty$.

  Bayesian interpretation: $y \sim N(X\beta, \sigma^2 I), \beta \sim N_p(0, \tau^2 I)$. $\hat{\beta}^{\text{ridge}}$ is the mean of $\beta|X,Y \sim N((X^TX+\frac{\sigma^2}{\tau^2}I)^{-1}X^TY, \Sigma')$. Here, $\lambda = \frac{\sigma^2}{\tau^2}$.

- Lasso

  Lasso ($\mathscr{L}_1$ regularization, least absolute shrinkage and selection operator): minimise $\frac{1}{2n}||y - x\beta||_2^2 + \lambda||\beta||_1 \Leftrightarrow$ minimise $||y - x\beta||_2^2$ s.t.$||\beta||_1 \le t$. Closed-form solution does not exist.

Variants: 1) Dantzig selector: minimise $||\beta||_1$ subject to $||\frac{1}{n}X^T(y-X\beta)||_\infty \le \lambda$.
Note: The Lasso solution satisfies the KKT(Karush-Kuhn-Tuckel) condition $\frac{1}{n}X^T(y-X\beta)+\lambda z=0$ where $z \in \partial||\cdot||_1=[-1,1]$. So $\frac{1}{n}||X^T(y-X\beta)||_\infty \le \lambda$.
2) Bridge: $\mathscr{L}_q(q\ge 1), \mathscr{L}_{1/2}$ regularization, etc.
3) SCAD: denote the regularization term as $\sum_{i=1}^p p_\lambda(\beta_j)$, then $p'_\lambda(t)=\lambda\text{sgn}(t)[I(t\le\lambda)+I(t>\lambda)\frac{(a\lambda-t)_+}{(a-1)\lambda}], a>2$. (release penalty for enough large λ)
4) MCP: $p'_\lambda(t)=\text{sgn}(t)\frac{(a\lambda-t)_+}{a}, a>1$. ("folded concave")
5) Elastic net: $\mathscr{L}_1+\mathscr{L}_2, \lambda\alpha||\beta||_1+\lambda(1-\alpha)||\beta||_2^2$. ("grouping effect")
6) Adaptive Lasso: $p_\lambda(\beta)=\lambda\sum_{j=1}^p w_j|\beta_j|, w_j=|\hat\beta_j|^{-\gamma}, \gamma>0$ where $\hat\beta_j$ is your initial estimate.

- Algorithms for Lasso
Prop: Under orthogonal design (i.e.$X^TX=I$), we have $\hat\beta^{\text{bs}}(\lambda)=\hat\beta^{\text{ls}}I(|\hat\beta^{\text{ls}}|\ge\lambda)$, $\hat\beta^{\text{ridge}}(\lambda)=\frac{1}{1+\lambda}\hat\beta^{\text{ls}}$, $\hat\beta^{\text{lasso}}(\lambda)=\text{sgn}(\hat\beta^{\text{ls}})(|\hat\beta^{\text{ls}}|-\lambda)_+$.
Alg 1: Least angle regresssion (LARS).
Idea: path following. Exploit the piecewise linearity of the solution path; homotopy algorithm. If a non-zero coefficient hits zero, drop its variable from the active set of variables and recompute the current joint least squares direction. Time complexity $O(p^3+p^2n)$, the same as least square, but low efficiency when p is large.
Alg 2: Coordinate descent, shooting algorithm.
With current estimate $\hat\beta$, minimising $f(\beta_j;\hat\beta)=\frac{1}{2}\sum_{i=1}^n(y_i-\sum_{k\ne j}x_{ik}\hat\beta_k-x_{ij}\beta_j)^2+\lambda\sum_{k\ne j}|\hat\beta_k|+\lambda|\beta_j|$ gives $\beta_j=S_\lambda(\sum_{i=1}^n x_{ij}(y_i-\sum_{k\ne j}x_{ik}\hat\beta_k))$ where $S_\lambda(x)=\text{sgn}(x)(|x|-\lambda)_+$ (after standardizing $||x_j||_2=1$). Cycle through $j=1,\cdots,p,1,\cdots,p,\cdots$.
Alg 3: Alternating direction method of multipliers (ADMM).
Idea: Split problem into two subproblems which can be solved alternatively. Minimise $f(\beta)+g(\gamma)$ s.t.$\beta-\gamma=0$ where $f(\beta)=\frac{1}{2}||y-x\beta||_2^2, g(\gamma)=\lambda||\gamma||_1$.
Augmented Lagrangian: $\mathscr{L}_\rho(\beta,\gamma,\alpha)=f(\beta)+g(\gamma)+(\tilde\alpha^T(\beta-\gamma)+\frac{\rho}{2}||\beta-\gamma||_2^2)$ (scaled form: $\frac{\rho}{2}||\beta-\gamma+\alpha||_2^2$ with $\alpha=\frac{\tilde\alpha}{\rho}$).
ADMM iterates $\beta^{k+1}\leftarrow(X^TX+\rho I)^{-1}(X^TY+\rho(\gamma^k-\alpha^k))$ (ridge), $\gamma^{k+1}\leftarrow S_{\lambda/\rho}(\beta^{k+1}+\alpha^k)$ (soft-thresholding), $\alpha^{k+1}\leftarrow\alpha^k+\beta^{k+1}-\gamma^{k+1}$.

- Theory for Lasso
Goals: 1) Variable selection consistency $P(\text{sgn}(\hat\beta_j)=\text{sgn}(\beta_j^*),\forall j)\ge 1-\delta$; 2) Nonasymptotic bounds on estimation/prediction $\forall$ finite p,n.
High-dimensional: ambient dim $p>>n$, but intrinsic dim $||\beta^*||_0=|\{j:\beta_k^*\ne 0\}|:=s<<n$.
Identifiability: when $p>n$, the Gram matrix $\Psi_n=\frac{1}{n}X^TX$ is singular, hence β^* is not identifiable.
Remedy: Assume Ψ_n is nondegenerate in certain "sparse" directions.
Restrained eigenvalue (RE) condition: $k(s,c_0)=\min_{J\subset\{1,\cdots,p\},|J|\le s}\min_{\delta\ne 0,||\delta_{J^c}||_1\le c_0||\delta_J||_1}\frac{||X\delta||_2}{\sqrt{n}||\delta_J||_2}>0$. (after standaridising $||x_j||_2=\sqrt{n}$) (RE(s,c_0) condition)
Lemma (Basic inequality): Let $S=\text{supp}(\beta^*), \epsilon\sim N(0,\sigma^2), \lambda=C\sigma\sqrt{\frac{\text{log}p}{n}}$ with $C>2\sqrt{2}$. Then with probability $\ge 1-p^{1-C^2/8}$, it holds that $\frac{1}{n}||X(\hat\beta-\beta^*)||_2^2+\lambda||\hat\beta-\beta^*||_1\le 4\lambda||\hat\beta_S-\beta_S^*||_1\le 4\lambda\sqrt{s}||\hat\beta_S-\beta_S^*||_2$.
Proof: By optimality of $\hat\beta$, $\frac{1}{2n}||Y-X\hat\beta||_2^2+\lambda||\hat\beta||_1\le\frac{1}{2n}||Y-X\beta^*||_2^2+\lambda||\beta^*||_1$. Substituting

$Y = X\beta^* + \epsilon$ gives $\frac{1}{2n}||\epsilon - X(\hat\beta - \beta^*)||_2^2 + \lambda||\hat\beta||_1 \le \frac{1}{2n}||\epsilon||_2^2 + \lambda||\beta^*||_1$ or $\frac{1}{2n}||X(\hat\beta - \beta^*)||_2^2 \le \frac{1}{n}\epsilon^T X(\hat\beta - \beta^*) + \lambda||\beta^*||_1 - \lambda||\hat\beta||_1$. If $\epsilon \sim N(0, \sigma^2)$, by Gaussian concentration inequality, $P(||\frac{1}{n}X^T\epsilon||_\infty > \frac{\lambda}{2}) \le \sum_{j=1}^p P(|\frac{1}{n}X_j^T\epsilon| > \frac{\lambda}{2}) \le pe^{-n\lambda^2/(8\sigma^2)}$. Take $\lambda = C\sigma\sqrt{\frac{\mathrm{log}p}{n}}$ with $C > 2\sqrt{2}$, then $pe^{-n\lambda^2/(8\sigma^2)} = p^{1-C^2/8} \to 0$. Conditional on the event $A = \left\{||\frac{1}{n}X^T\epsilon||_\infty \le \frac{\lambda}{2}\right\}$ which holds with probability $\ge 1 - p^{1-C^2/8}$, $\frac{1}{n}||X(\hat\beta - \beta^*)||_2^2 \le \lambda||\hat\beta - \beta^*||_1 + 2\lambda(||\beta^*||_1 - ||\hat\beta||_1)$. Adding $\lambda||\hat\beta - \beta^*||_2^2$ to both sides, $\frac{1}{n}||X(\hat\beta - \beta^*)||_2^2 + \lambda||\hat\beta - \beta^*||_1 \le 2\lambda(||\hat\beta - \beta^*||_1 + ||\beta^*||_1 - ||\hat\beta||_1) = 2\lambda(||\hat\beta_S - \beta_S^*||_1 + ||\beta_S^*||_1 - ||\hat\beta_S||) \le 4\lambda||\hat\beta_S - \beta_S^*||_1$.

Corollary: On event A, $||\delta_{S^c}||_1 \le 3||\delta_S||_1$ where $\delta = \hat\beta - \beta^*$.

Proof: Basic inequality implies $\lambda||\delta||_1 \le 4\lambda||\delta_S||_1$, or $||\delta_{S^c}||_1 \le 3||\delta_S||_1$.

Thm (nonasymptotic bound): $||\hat\beta - \beta^*||_1 \le \frac{16C}{k^2}\sigma s\sqrt{\frac{\mathrm{log}p}{n}}$ with probability $\ge p^{1-C^2/8}$.

Proof: By the basic inequality, $\frac{1}{n}||X\delta||_2^2 \le 4\lambda\sqrt{s}||\delta_S||_2$. On the other hand, by RE$(s, 3)$, $\frac{||X\delta||_2}{\sqrt{n}||\delta_S||_2} \ge k > 0$ or $||\delta_S||_2 \le \frac{||X\delta||_2}{\sqrt{n}k}$. Combining $\frac{1}{n}||X\delta||_2^2 \le 4\lambda\sqrt{s}\frac{||X\delta||_2}{\sqrt{nk}}$ or $\frac{1}{n}||X\delta||_2^2 \le \frac{16\lambda s}{k^2} = \frac{16C^2}{k^2}\sigma^2\frac{s\mathrm{log}p}{n}$. Moreover, $||\delta||_1 \le 4||\delta_S||_1 \le 4\sqrt{s}||\delta_S||_2 \le \frac{4\sqrt{s}}{k}\frac{||X\delta||_2}{\sqrt{n}} \le \frac{4\sqrt{s}}{k}\frac{4C}{k}\sigma\sqrt{\frac{s\mathrm{log}p}{n}} = \frac{16C}{k^2}\sigma s\sqrt{\frac{\mathrm{log}p}{n}}$.

Prediction consistency: l_2 estimate $s\mathrm{log}p = o(n)$ and l_1 estimate $s^2\mathrm{log}p = o(n)$.

Q: Is the Lasso still prediction consistent without assumption on X?

Thm: Assume $||\beta^*||_1 \le K$. Then solution $\hat\beta$ to minmise$_\beta||Y - X\beta||_2^2$ s.t.$||\beta||_1 \le K$ satisfies with high probability, $\frac{1}{n}||X(\hat\beta - \beta^*)||_2^2 \le CK\sigma\sqrt{\frac{\mathrm{log}p}{n}}$.

Proof: By defnition, $\hat Y$ is the projection of Y onto the compact convex set $\mathcal{C} = \{X\beta : ||\beta||_1 \le K\}$. Also, since $||\beta^*||_1 \le K$ we have $Y^* = X\beta^* \in \mathcal{C}$. Thus, $0 \ge (\hat Y - Y^*)^T(\hat Y - Y) = (\hat Y - Y^*)^T(\hat Y - Y^* - Y + Y^*) = ||\hat Y - Y^*||_2^2 - (\hat Y - Y^*)^T(Y - Y^*)$ or $\frac{1}{n}||\hat Y - Y^*||_2^2 \le \frac{1}{n}(\hat Y - Y^*)^T(Y - Y^*) = \frac{1}{n}\epsilon^T X(\hat\beta - \beta^*) \le ||\frac{1}{n}\epsilon^T X||_\infty||\hat\beta - \beta^*||_1$ (on A) $\le \frac{\lambda}{2}2K = CK\sigma\sqrt{\frac{\mathrm{log}p}{n}}$.

- Extensions

  Group Lasso: $y = \sum_{g=1}^G x_g\beta_g + \epsilon$ where $x_g \in M_{n\times p_g}(R)$ and $\beta_g \in M_{p_g\times 1}(R)$. Penalty: $p_\lambda(\beta) = \lambda\sum_{g=1}^G \sqrt{p_g}||\beta_g||_2$.

  Example: Multivariate regression: $Y_{n\times q} = X_{n\times p}B_{p\times q} + E$.

  Lasso (entrywise): $\lambda||B||_1 = \lambda\sum_{i,j}|b_{ij}|$.

  Prediction selection: select predictors having effects on all responses.

  Reduced rank regression (RRR): minimise $||Y - XB||_F^2$ or $\mathrm{tr}((Y - XB)^T(Y - XB)\Sigma^{-1})$ s.t. $\mathrm{rank}(B) \le r$ Denote non-zero eigenvalue of $(B^TB)^{\frac{1}{2}}$ as $\sigma_j(B)$, and $\mathrm{rank}(B) = ||\sigma(B)||_0$. Also we have a new regression method: minimise$||Y - XB||_F^2 + \lambda||B||_*$, where $||B||_*$ is nuclear norm: $\sum_{j=1}^r \sigma_j(B) = ||\sigma(B)||_1 = \mathrm{tr}((B^TB)^{\frac{1}{2}})$. (the last Ky Fan norm)

  Sparse LDA: Idea: Exploit connection between LDA nad LS.

  Lemma: Label the classes $y_1 = -\frac{n}{n_1}, y_2 = \frac{n}{n_2}$, where $n = n_1 + n_2$. Let $\hat\beta^{\mathrm{ls}} = \mathrm{argmin}_\beta\sum_{i=1}^n(y_i - \beta_0 - x_i^T\beta)^2$, then $\hat\beta^{\mathrm{ls}} = C\hat\Sigma^{-1}(\hat\mu_2 - \hat\mu_1)$ for some $C > 0$.

  $\mathscr{L}_1$ regularization version: minimise$_\beta\sum_{i=1}^n(y_i - \beta_0 - x_i^T\beta)^2 + \lambda||\beta||_1$.

  Dantzig selection type: minimise $||\beta||_1$ s.t. $||\hat\Sigma\beta - (\hat\mu_1 - \hat\mu_2)||_\infty < \lambda$.

  Sparse GLMs (Logistic, Possion, etc.): minimise$_\beta - \mathrm{loglik}(\beta) + \lambda||\beta||_1$.

6 Support Vector Machines

- Separable case

  Separality hyperplanes: $L : \beta_0 + \beta^T x = 0$. $\forall x_1, x_2 \in L, \beta^T(x_1 - x_2) = 0; \forall x_0 \in L, \beta^T x_0 = -\beta_0$. Denote $\beta^* = \frac{\beta}{||\beta||}$. Signed distance: $\beta^{*T}(x - x_0) = \frac{1}{||\beta||}(\beta^T x - \beta^T x_0) = \frac{1}{||\beta||}(\beta^T x + \beta_0)$.

  Perception learning: $\hat{f}(x) = \hat{\beta}^T x + \hat{\beta}_0, \hat{G}(x) = \text{sgn}(\hat{f}(x))$. minimise $D(\beta, \beta_0) = -\sum_{i\in M} y_i(x_i^T\beta + \beta_0)$ where M is misclassified set.

  Algorithm: $\frac{\partial D}{\partial \beta} = -\sum_{i\in M} y_i x_i, \frac{\partial}{\beta_0} = -\sum_{i\in M} y_i$. Randomly pick a misclassified point x_i and update $(\beta, \beta_0)^T \leftarrow (\beta, \beta_0)^T + \rho(y_i x_i, y_i)^T$ where ρ is learning rate. (SGD)

- Nonseparable case

  Goals: maximise$_{\beta,\beta_0,||\beta||=1} M$ s.t.$y_i(x_i^T\beta + \beta_0) \geq M, \forall i$. Let $M = \frac{1}{||\beta||}$, rewrite as minimise$_{\beta,\beta_0} ||\beta||$ s.t.$y_i(x_i^T\beta + \beta_0) \geq 1$.

  Introducing slack variables $\xi_1, \cdots, \xi_n$, s.t.$y_i(x_i^T\beta + \beta_0) \geq M(1 - \xi_i), \xi_i \geq 0, \sum_{i=1}^n \xi_i \leq K$.

  Optimization problem: $\min_{\beta,\beta_0,\xi} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^n \xi_i$ s.t. $\xi_i \geq 0, y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, \forall i$.

  Lagrangian: $\mathscr{L}(\beta, \beta_0, \xi, \alpha, \mu) = \frac{1}{2}||\beta||^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i[y_i(x_i^T\beta + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^n \mu_i\xi_i$.

  Setting derivatives to zero, $\frac{\partial \mathscr{L}}{\partial \beta} = \beta - \sum_{i=1}^n \alpha_i y_i x_i = 0, \frac{\partial \mathscr{L}}{\partial \beta_0} = -\sum_{i=1}^n \alpha_i y_i = 0, \frac{\partial \mathscr{L}}{\partial \xi_i} = C - \alpha_i - \mu_i = 0$. In addition, $\alpha_i, \mu_i, \xi_i \geq 0$. Dual problem: $\max_{\beta,\beta_0,\xi,\alpha,\mu} \mathscr{L}(\cdot)$, s.t.$\nabla\mathscr{L} = 0, \alpha_i, \mu_i \geq 0$. Dual objective function: $\mathscr{L}_D = \frac{1}{2}||\sum_{i=1}^n \alpha_i y_i x_i||^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i[y_i(x_i^T \sum_j \alpha_j y_j x_j + \beta_0) - 1 + \xi_i] - \sum_i \mu_i\xi_i = \sum_i \alpha_i - \frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j x_i^T x_j$ s.t. $0 \leq \alpha_i \leq C, \sum_{i=1}^N \alpha_i y_i = 0$.

  Why called SVM? Complementary slackness: $\alpha_i[y_i(x_i^T\beta + \beta_0) - 1 + \xi_i] = 0, \mu_i\xi_i = 0, \forall i$. Because $\hat{\beta} = \sum_{i=1}^n \hat{\alpha}_i y_i x_i$ and $\alpha_i \neq 0$ iff $y_i(x_i^T\beta + \beta_0) = 1 - \xi_i$, $\hat{\beta}$ is only determined by support vector. 1) $\hat{\xi}_i = 0$, on the margin's boundary, or $0 < \hat{\alpha}_i < C$; 2) $\hat{\xi}_i = 0$, or $\hat{\alpha}_i = C$.

- Kernel method: nonlinear boundaries

  Transfrom the input into $h(x_i) = (h_1(x_i), \cdots, h_B(x_i))$.

  Lagrange dual: $\mathscr{L}_D = \sum_i \alpha_i - \frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j (h(x_i), h(x_j))$.

  Reproducing kernel Hilbert spaces: Idea: Generalize covariance to infinite dimensions.

  Def: A bivariate function K on $E \times E$ is called a reproducing kernel for a Hilbert space $\mathscr{H}$ if a) $\forall t \in E, K(\cdot, t) \in \mathscr{H}$; b) (reproducing property) $\forall f \in \mathscr{H}$ and $\forall t \in E, f(t) = (f, K(\cdot, t))$. When such a RK exists, $\mathscr{H}$ is called a RKHS.

  Example: $\mathscr{H}$ is finite-dim, with orthogonal basis $\{e_1, \cdots, e_p\}$. Define $K(s, t) = \sum_{i=1}^p e_i(s)e_i(t)$. Check: a) $K(\cdot, t) = \sum_{i=1}^p e_i(\cdot)e_i(t) \in \mathscr{H}$; b) $(e_j, K(\cdot, t)) = (e_j, \sum_{i=1}^p e_i(\cdot)e_i(t)) = e_j(t)$.

  Def: K on $E \times E$ is positive semi-definite(PSD) if $\forall n$,$\{t_1, \cdots, t_n\} \subset E, \sum_{i=1}^n \sum_{j=1}^n a_i a_j K(t_i, t_j) \geq 0$.

  Thm: The RK K of a RKHS $\mathscr{H}$ is unique, symmetric and PSD.

  Proof: Suppose $\exists K_1, K_2$ for $\mathscr{H}$. Then $f(t) = (f, K_1(\cdot, t)) = (f, K_2(\cdot, t))$ or $(f, (K_1 - K_2)(\cdot, t)) = 0, \forall f, \forall t \Rightarrow ((K_1 - K_2)(\cdot, t), (K_1 - K_2)(\cdot, t)) = 0, \forall t \in E \Rightarrow K_1 = K_2$.

  Symmetry: $K(s, t) = (K(\cdot, t), K(\cdot, s)) = (K(\cdot, s), K(\cdot, t)) = K(t, s)$.

  PSD: $\sum_i\sum_j a_i a_j K(t_i, t_j) = \sum_i\sum_j a_i a_j (K(\cdot, t_i), K(\cdot, t_j)) = (\sum_i a_i K(\cdot, t_i), \sum_j a_j K(\cdot, t_j)) \geq 0$.

  Thm: $\forall$ symmetric PSD function $K(\cdot, \cdot)$, $\exists$ a unique RKHS.

  Consider the regularized problem in RKHS: minimise$_{f\in\mathscr{H}} \sum_{i=1}^n \mathcal{L}(y_i, f(x_i)) + \lambda J(f)$ where $\lambda J(f)$ is penalty function. Orthogonal basis $\{\phi_i\}_{i=1}^\infty : f(x) = \sum_{i=1}^\infty c_i\phi_i(x), K(x, y) = \sum_{i=1}^\infty \gamma_i\phi_i(x)\phi_i(y)$.

Let $J(f)=||f||^2_{\mathscr{H}}$, we get generalized ridge: minimise$_{\{c_j\}_{j=1}^\infty}\sum_{i=1}^n\mathcal{L}(y_i,\sum_{j=1}^\infty c_j\phi_j(x_i))+\lambda||f||^2_{\mathscr{H}}$.
Let $\tilde K=(K(x_i),K(x_j))_{i,j=1}^n, f(x)=\sum_{i=1}^n\alpha_iK(x,x_i),\tilde\alpha=(\alpha_1,\cdots,\alpha_n)^T$. Then $J(f)=||f||^2_{\mathscr{H}}=(\sum_{i=1}^n\alpha_iK(\cdot,x_i),\sum_{j=1}^n\alpha_jK(\cdot,x_j))=\sum_{i=1}^n\sum_{j=1}^n\alpha_i\alpha_jK(x_i,x_j)=\tilde\alpha^T\tilde K\tilde\alpha$. So we transfer original problem to minmisize$_{\tilde\alpha}\mathcal{L}(\tilde y,\tilde K\tilde\alpha)+\lambda\tilde\alpha^T\tilde K\tilde\alpha$.
Method: 1) Kernel smoothing; 2) Splines, basis spansion (wavelets, etc.).
Choices of kernel function: 1) polynomial: $(1+(x_i,x_j))^\alpha$; 2) radial basis: $e^{-\gamma||x_i-x_j||^2}$; 3) Sigmoid: $\tanh(k_1(x_i,x_j)+k_2)$.

- Regularization on SVMs & SV regression
Our goal is to $\min_{\beta,\beta_0,\xi}\frac{1}{2}||\beta||^2+C\sum_{i=1}^n\xi_i$ s.t.$\xi_i\geq 0,y_i(x_i^T\beta+\beta_0)\geq 1-\xi_i,\forall i\Leftrightarrow\min_{\beta,\beta_0}\frac{1}{2}||\beta||^2+C\sum_{i=1}^n(1-y_if(x_i))_+$ or $\min_{\beta,\beta_0}\sum_{i=1}^n(1-y_if(x_i))_++\frac{\lambda}{2}||\beta||^2$ ($\lambda=\frac{1}{C}$) := hinge loss + regularization.
C large, λ small: ξ_i small, $||\beta||$ large, wiggly boundary, tends to overfit.
C small, λ large: ξ_i large, $||\beta||$ small, smooth boundart, tends to underfit.
Two key ingredients for SVMs: 1) hinge loss (soft margin); 2) kernel trick (dual problem is simple).
Nonparametric setting: Suppose K has the eigen-expansion: $K(x,y)=\sum_{m=1}^\infty\sigma_m\phi_m(x)\phi_m(y)$, then $h_m(x)=\sqrt{\sigma}\phi_m(x)$. Optimization problem becomes min $\sum_{i=1}^n(1-y_i(\beta_0+\sum_{m=1}^\infty\theta_m\phi_m(x_i)))_++\frac{\lambda}{2}\sum_{m=1}^\infty\frac{\theta_m^2}{\sigma_m}$ where $\theta_m=\sqrt{\sigma_m}\beta_m$. We guess ∃ a finite-dim solution: $f(x)=\beta_0+\sum_{i=1}^n\alpha_iK(x,x_i)$, where $\beta_0,\alpha=\text{argmin}_{\beta_0,\alpha}\sum_{i=1}^n(1-y_if(x_i))_++\frac{\lambda}{2}\tilde\alpha^T\tilde K\tilde\alpha$.
Regression: replace hinge loss by $\mathcal{L}_\epsilon(y,f)=(|y-f|-\epsilon)_+$, our goal is to $\min_{\beta,\beta_0}\mathcal{L}_\epsilon(y_i,f(x_i))+\frac{\lambda}{2}||\beta||^2$, where $f(x)=x^T\beta+\beta_0$.
Introduce slack variables $\xi_i\eta_i$, the goal turns to $\min_{\beta,\beta_0,\xi_i,\eta_i}\sum_{i=1}^n(\xi_i+\eta_i)+\frac{\lambda}{2}||\beta||^2$ s.t. $y_i-f(x_i)\geq-\epsilon-\xi_i,y_i-f(x_i)\leq\epsilon+\eta_i,\xi_i\geq 0,\eta_i\geq 0,\forall i$.

- Margin Theory
Using VCdim-based generalization bound, $\mathcal{R}(h)\leq\hat{\mathcal{R}}_S(h)+\sqrt{\frac{2d\log(em/d)}{m}}+\sqrt{\frac{\log(1/\delta)}{2m}}$. VCdim$(R^p)=p+1$, $\mathcal{R}(h)\leq\hat{\mathcal{R}}_S(h)+O(\sqrt{\frac{\log(m/p+1)}{m/(p+1)}})$, which requires $p<<m$.
Goal: Dim-free, margin-based generalization bounds.
Def: Confidence margin: $yh(x)$, ρ-margin loss $\Phi_\rho(x)=1\,(x\leq 1),\ 1-\frac{x}{\rho}\,(0<x<\rho),\ 0\,(x\geq\rho)$, empirical $\hat{\mathcal{R}}_{S,\rho}(h)=\frac{1}{m}\sum_{i=1}^m\Phi_\rho(y_ih(x_i))\leq\frac{1}{m}\sum_{i=1}^m I(y_ih(x_i)\leq\rho)$.
Lemma(Talagrand): If Φ in L-Lipschitz, then ∀ hypothesis set $\mathcal{H}$, $\hat{\text{Rad}}_S(\Phi\circ\mathcal{H})\leq L\hat{\text{Rad}}_S(\mathcal{H})$.
Thm: Fix $\rho>0$, $\forall\delta>0$, the following holds $\forall h\in\mathcal{H}$, with prob $\geq 1-\delta$, $\mathcal{R}(h)\leq\hat{\mathcal{R}}_{S,\rho}(h)+\frac{2}{\rho}\text{Rad}_m(\mathcal{H})+\sqrt{\frac{\log(1/\delta)}{2m}}$.
Proof: Let $G=\{z=(x,y)\to yh(x):h\in\mathcal{H}\}$. By Rad generalization bound, $\forall h\in\mathcal{H}$, with prob $\geq 1-\delta$, $E\Phi_\rho(yh(x))\leq\hat{\mathcal{R}}_{S,\rho}+2\text{Rad}_m(\Phi_\rho\circ G)+\sqrt{\frac{\log(1/\delta)}{2m}}$. Since $\mathcal{R}(h)=EI(yh(x)\leq 0)\leq E\Phi_\rho(yh(x))$, & by above lemma, $\text{Rad}_m(\Phi_\rho\circ G)\leq\frac{1}{\rho}\text{Rad}_m(G)=\frac{1}{\rho m}E_{S,\sigma}\text{sup}_{h\in\mathcal{H}}\sum_{i=1}^m\sigma_iy_ih(x_i)=\frac{1}{\rho}\text{Rad}_m(\mathcal{H})$.
Remark: This can be made uniform over $\forall\rho\in[0,r]$ where r is fixed: $\mathcal{R}(h)\leq\hat{\mathcal{R}}_{S,\rho}(h)+\frac{4}{\rho}\text{Rad}_m(\mathcal{H})+\sqrt{\frac{\log\log_2(2r/\rho)}{m}}+\sqrt{\frac{\log(2/\delta)}{2m}}$.
Thm: Let $S\subset\{x:||x||\leq r\}$ and $\mathcal{H}=\{x\to(w,x):||w||\leq B\}$, then $\hat{\text{Rad}}_S(\mathcal{H})\leq\frac{Br}{\sqrt{m}}$.
Proof: $\hat{\text{Rad}}_S(\mathcal{H})=\frac{1}{m}E_\sigma\text{sup}_{||w||\leq B}\sum_{i=1}^m\sigma_i(w,x_i)=\frac{1}{m}E_\sigma\text{sup}_{||w||\leq B}(w,\sum_{i=1}^m\sigma_ix_i)\leq\frac{B}{m}E_\sigma||\sum_{i=1}^m\sigma_i x_i||\leq\frac{B}{m}\sqrt{E_\sigma||\sum_{i=1}^m\sigma_ix_i||^2}\leq\frac{B}{m}\sqrt{\sum_{i=1}^m||x_i||^2}\leq\frac{Br}{\sqrt{m}}$.
Corollary: For fixed $\rho>0$, $\mathcal{R}(h)\leq\hat{\mathcal{R}}_{S,\rho}(h)+\frac{2Br}{\rho\sqrt{m}}+3\sqrt{\frac{\log(2/\delta)}{2m}}$.

7 Boosting

- Idea: Sequentially combine a set of weak learners into a single strong learner.
  Weak classifiers $G_m(x), m = 1, \cdots, M$. Combine to form $G(x) = \text{sgn}(\sum_{m=1}^{M} \alpha_m G_m(x))$.

- AdaBoost
  1) Initialize $w_i = \frac{1}{n}, i = 1, \cdots, n$.
  2) For $m = 1, \cdots, M$:
  —— a) fit $G_m(x)$ to training data weighted by w_i;
  —— b) $\epsilon_m = \frac{\sum w_i I(y_i \neq G_m(x_i))}{\sum w_i}$;
  —— c) $\alpha_m = \log \frac{1-\epsilon_m}{\epsilon_m}$;
  —— d) update weights to overrepresent misclassified cases: $w_i \leftarrow w_i \exp(\alpha_m I(y_i \neq G_m(x_i)))$.
  3) $G(x) = \text{sgn}(\sum \alpha_m G_m(x))$.

- Statistical view
  Additive models: $y_i = \sum_{j=1}^{J} \beta_j f(x_j; \gamma_j) + \epsilon$.
  Forward stagewise additive modeling:
  1) Initialize $f_0(x) = 0$;
  2) For $m = 1, \cdots, M$:
  —— a) $(\beta_m, \gamma_m) = \text{argmin}_{\beta,\gamma} \sum_{i=1}^{n} \mathcal{L}(y_i, f_{m-1}(x_i) + \beta G_\gamma(x_i))$;
  —— b) $f_m(x) = f_{m-1}(x) + \beta_m G_{\gamma_m}(x)$.
  Take expotential loss: $\mathcal{L}(y, f(x)) = e^{-yf(x)}$, then $(\beta_m, G_m) = \text{argmin}_{\beta,G} \sum_{i=1}^{n} \exp(-y_i(f_{m-1}(x) + \beta G(x_i))) = \sum_{i=1}^{n} e^{-y_i f_{m-1}(x_i)} e^{-\beta y_i G(x_i)} (w_i^{(m)} = e^{-y_i f_{m-1}(x_i)}) = e^{-\beta} \sum_{i:y_i=G(x_i)} w_i^{(m)} + e^{\beta} \sum_{i:y_i \neq G(x_i)} w_i^{(m)} = (e^{\beta} - e^{-\beta}) \sum_{i=1}^{m} w_i^{(m)} I(y_i \neq G(x_i)) + e^{-\beta} \sum_{i=1}^{m} w_i^{(m)}$. So that $G_m = \text{argmin}_G \sum_{i=1}^{m} w_i^{(m)} I(y_i \neq G(x_i)), \beta_m = \text{argmin}_\beta (e^{\beta} - e^{-\beta})\epsilon_m + e^{-\beta} = \text{argmin}_\beta e^{\beta}\epsilon_m + e^{-\beta}(1-\epsilon_m) = \frac{1}{2}\log\frac{1-\epsilon_m}{\epsilon_m}$. By the define of $w_i^{(m)}$, $w_i^{(m+1)} = w_i^{(m)} e^{-\beta_m y_i G_m(x_i)} = w_i^{(m)} e^{2\beta_m I(y_i \neq G_m(x_i))} e^{-\beta_m} = w_i^{(m)} e^{\alpha_m I(y_i \neq G_m(x_i))} e^{-\beta_m}$.
  Remark: $f(x) = \text{argmin}_{f(x)} E_y e^{-yf(x)} = \frac{1}{2}\log\frac{P(y=1|x)}{P(y=-1|x)}$.

- Loss functions and robustification
  a) misclassification error: $I(y \neq \text{sgn}(f))$;
  b) binomial deviance: $\log(1 + e^{-2yf})$;
  c) square error: $(y - f)^2$;
  d) hinge loss: $(1 - yf)_+$;
  e) expotential loss: e^{-yf}.

- Boosting for regression
  a) squared loss: $L(y, f(x)) = (y - f(x))^2$.
  b) absolute loss: $L(y, f(x)) = |y - f(x)|$.
  c) Huber loss: $L(y, f(x)) = \begin{cases} [y - f(x)]^2, & |y - f(x)| \leq \delta \\ 2\delta(|y - f(x)| - \delta^2), & \text{otherwise} \end{cases}$.

- Gradient boosting
  Choose $h_m = -\rho_m g_m$ where ρ_m is a scalar and $g_m \in R^n$.

$g_{mi} = \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}|_{f(x_i)=f_{m-1}(x_i)}$ and $\rho_m = \text{argmin}_\rho L(f_{m-1} - \rho g(m))$.
The above is only defined at the traning data points x_i, thus we can induce a tree $T(x; \Theta_m)$ at the m-th iteration whose predictions t_m are as close as possible to the negative gradient. Using squared error: $\hat{\Theta}_m = \text{argmin}_\Theta \sum_{i=1}^N (-g_{mi} - T(x_i, \Theta))^2$.

- Theory for AdaBoost
  Assume sample size m, base classifiers h_t and T rounds of boosting.
  Thm: The empirical error of AdaBoost $\hat{\mathcal{R}}_S(f) \le \exp(-2\sum_{t=1}^T (\frac{1}{2} - \epsilon_t)^2)$. Moreover, if $\gamma \le \frac{1}{2} - \epsilon_t, \forall t$, then $\hat{\mathcal{R}}_S(f) \le e^{-2\gamma^2 T}$.
  Proof: $\hat{\mathcal{R}}_S(f) = \frac{1}{m}\sum_{i=1}^m I(y_i f(x_i) \le 0) \le \frac{1}{m}\sum_{i=1}^m e^{-y_i f(x_i)}$. Let Z_t be the normalization factor of distribution weights, $w_i^{(T+1)} = \frac{1}{Z_i} w_i^{(T)} e^{-\alpha_T y_i h_T(x_i)} = \frac{1}{Z_{T-1}Z_T} w_i^{(T-1)} e^{-\alpha_T y_i h_T(x_i)} e^{-\alpha_{T-1} y_i h_{T-1}(x_i)} = \cdots = \frac{1}{m\prod_{t=1}^T Z_t} \exp(-y_i \sum_{t=1}^T \alpha_t h_t(x_i))$. Therefore $\hat{\mathcal{R}}_S(f) \le \prod_{t=1}^T Z_t \sum_{i=1}^m w_i^{(T+1)} = \prod_{t=1}^T Z_t$. Also, $Z_t = \sum_{i=1}^m w_i^{(t)} e^{-\alpha_t y_i h_t(x_i)} = \sum_{i:y_i = h_t(x_i)} w_i^{(t)} e^{-\alpha_t} + \sum_{i:y_i \ne h_t(x_i)} w_i^{(t)} e^{\alpha_t} = (1-\epsilon_t)e^{-\alpha_t} + \epsilon_t e^{\alpha_t} = (1-\epsilon_t)\sqrt{\frac{\epsilon_t}{1-\epsilon_t}} + \epsilon_t \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} = 2\sqrt{\epsilon_t(1-\epsilon_t)}$. Thus, $\prod_{t=1}^T Z_t = \prod_{t=1^T} 2\sqrt{\epsilon_t(1-\epsilon_t)} = \prod_{t=1}^T \sqrt{1 - 4(\frac{1}{2} - \epsilon_t)^2} \le \prod_{t=1}^T \exp(-2(\frac{1}{2} - \epsilon_t)^2) = \exp(-2\sum_{t=1}^T (\frac{1}{2} - \epsilon_t)^2)$.
  Remark: “Adaptive” to γ or ϵ_t.

- Margin theory
  Hypothesis set for AdaBoost: $F_T = \{\text{sgn}(\sum_{t=1}^T \alpha_t h_t) : \alpha_t \ge 0, h_t \in \mathcal{H}\}$.
  $\text{VCdim}(F_T) \le 2(d+1)(T+1)\log_2((T+1)e) = O(dT\log T)$ where $d = \text{VCdim}(\mathcal{H})$.
  Not useful for large T. The reality is test error $\equiv 0$ but generalization error $\downarrow$ with $T \uparrow$).
  Def: $\mathscr{L}_1$ (geometric) margin $\rho_f(x) = \frac{|f(x)|}{||\alpha||_1} = \frac{|(\alpha, h(x))|}{||\alpha||_1}$, $\rho_f = \min_{1 \le i \le m} \rho_f(x_i)$. This is the confidence margin of $\bar{f} = \frac{f}{||\alpha||_1}$.
  Let $\text{conv}(\mathcal{H}) = \{\sum_{j=1}^p \mu_j h_j : p \ge 1, \mu_j \ge 0, h_j \in \mathcal{H}, \sum_{j=1}^p \mu_j \le 1\}$ be the convex hull of $\mathcal{H}$.
  Lemma: $\hat{\text{Rad}}_S(\text{conv}(\mathcal{H})) = \hat{\text{Rad}}_S(\mathcal{H})$.
  Proof: $\hat{\text{Rad}}_S(\text{conv}(\mathcal{H})) = \frac{1}{m} E_\sigma \sup_{h_j \in \mathcal{H}, \mu_j \ge 0, ||\mu||_1 \le 1} \sum_{i=1}^m \sigma_i \sum_{j=1}^p \mu_j h_j(x_i)$
  $= \frac{1}{m} E_\sigma \sup_{h_j \in \mathcal{H}} \sup_{\mu_j \ge 0, ||\mu||_1 \le 1} \sum_{j=1}^p \mu_j \sum_{i=1}^m \sigma_i h_j(x_i) = \frac{1}{m} E_\sigma \sup_{h_j \in \mathcal{H}} \max_{1 \le j \le p} \sum_{i=1}^m \sigma_i h_j(x_i)$
  $= \frac{1}{m} E_\sigma \sup_{h \in \mathcal{H}} \sum_{i=1}^m \sigma_i h(x_i) = \hat{\text{Rad}}_S(\mathcal{H})$.
  Corollary: $\mathcal{R}(f) \le \hat{\mathcal{R}}_{S,\rho}(\bar{f}) + \frac{2}{\rho}\text{Rad}_m(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2m}}$ with prob at least $1 - \delta$.
  Finally, we verify that $\hat{\mathcal{R}}_{S,\rho}(\bar{f})$ decays exponentially with T.
  Thm: $\hat{\mathcal{R}}_{S,\rho}(\bar{f}) \le 2^T \prod_{t=1}^T \epsilon_t^{(1-\rho)/2}(1-\epsilon_t)^{(1+\rho)/2}$.
  Proof: $\hat{\mathcal{R}}_{S,\rho}(\bar{f}) \le \frac{1}{m}\sum_{i=1}^m I(\frac{y_i f(x_i)}{||\alpha||_1} \le \rho) \le \frac{1}{m}\sum_{i=1}^m e^{-y_i f(x_i) + \rho||\alpha||_1} = e^{\rho||\alpha||_1} \prod_{t=1}^T Z_t$
  $= \exp(\frac{\rho}{2}\sum_{t=1}^T \log\frac{1-\epsilon_t}{\epsilon_t}) \prod_{t=1}^T 2\sqrt{\epsilon_t(1-\epsilon_t)} = 2^T \prod_{t=1}^T \epsilon_t^{(1-\rho)/2}(1-\epsilon)^{(1+\rho)/2}$.
  Remark: If $\gamma \le \frac{1}{2} - \epsilon_t$ and $\rho \le 2\gamma$, then the upper bound is maximized at $\epsilon_t = \frac{1}{2} - \gamma$: $\hat{\mathcal{R}}_{S,\rho}(\bar{f}) \le [(1-2\gamma)^{1-\rho}(1+2\gamma)^{1+\rho}]^{T/2} = [(1-4\gamma^2)(\frac{1+2\gamma}{1-2\gamma})^\rho]^{T/2} < 1$ when $\rho < \gamma$.

- Regularization for AdaBoost
  Idea: Prevent the algorithm from concentrating on a few base learners and/or hard examples.
  Method 1: Early stopping.
  Method 2: $\mathscr{L}_1$ regularization: $\text{minimise}_{\alpha \ge 0} \frac{1}{m}\sum_{i=1}^m e^{-y_i f(x_i)} + \lambda||\alpha||_1$. Why?
  By margin bounds, $\forall f = \sum_j \alpha_j h_j$ with $||\alpha||_1 \le 1$, $\mathcal{R}(f) \le \frac{1}{m}\sum_{i=1}^m e^{1 - y_i f(x_i)/\rho} + \frac{2}{\rho}\text{Rad}_m(\mathcal{H}) + \cdots$.
  or, since f/ρ has the same generalization error as f, $\forall ||\alpha||_1 \le 1/\rho, \mathcal{R}(f) \le \frac{1}{m}\sum_{i=1}^m e^{1 - y_i f(x_i)} + \cdots$.

8 Clustering and Dimension Reduction

- Clustering

  Dissimilarity measure $D = (d_{ij})$. Given clusters $K < n$, dertermine cluster assignmeng $k = C(i)$.
  Total point scatter $T = \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n d_{ij} = \frac{1}{2}\sum_{k=1}^K \sum_{C(i)=k}(\sum_{C(j)=k} d_{ij} + \sum_{C(j)\neq k} d_{ij}) := W(C) + B(C) =$ within-cluster + between cluster.
  Goal: find the optimal C^* minimizing $W(C)$ or equivalently, maximizing $B(C)$.
  Greedy algorithm: Use Euclidean distance as dissimilarity measure: $d_{ij} = ||x_i - x_j||^2$. $W(C) = \frac{1}{2}\sum_{k=1}^K \sum_{C(i)=k}\sum_{C(j)=k} ||x_i - x_j||^2 = \sum_{k=1}^K n_k \sum_{C(i)=k} ||x_i - \bar{x}_k||^2$ where $\bar{x}_k = \frac{1}{n_k}\sum_{C(i)=k} x_i$ and $n_k = |\{i : C(i) = k\}|$.
  K-means: 1) Fix cluster assignment C, $\text{minmisize}_{m_k} \sum_{C(i)=k} ||x_i - m_k||^2$;
  2) Fix means m_k, update cluster assignment $C(i) = \text{argmin}_k ||x_i - m_k||^2$.
  Remarks: a) Desent property; b) Multiple random starts.
  Soft K-means (Gaussian mixture generative model): $g(x) = \sum_{k=1}^K \pi_k g_k(x)$, g_k pdf of $\mathcal{N}(\mu_k, \sigma^2)$, $\pi_k \geq 0$, $\sum_k \pi_k = 1$.
  EM algorithm: 1) E-step: compute soft assignment $\hat{\gamma}_{ik} = \frac{\hat{\pi}_k g_k(x_i;\hat{\mu}_k,\hat{\sigma}^2)}{\sum_l \hat{\pi}_l g_l(x_i;\hat{\mu}_l,\hat{\sigma}^2)}$;
  2) M-step: update weighted means and variances.
  Spectral clustering: 1) connected components; 2) not locally clustered.
  Idea: Use tools of graph theory to reduce dimension and apply K-means to the transformed data.
  Weighted undirected graph(network) $G = (V, Z)$, adjacency matrix $W = (w_{ij})_{n\times n}, w_{ij} \geq 0$, e.g.$w_{ij} = \exp(-d_{ij}^2/\gamma)$, KNN, etc.
  Graph Laplacian: $L = D - W$ where $D = \text{diag}(d_1, \cdots, d_n), d_i = \sum_j w_{ij}$.
  Normalized Laplacian: $\widetilde{L} = D^{-1/2}LD^{-1/2} = I - D^{-1/2}WD^{-1/2}$.
  Prop: L satisfies a) $\forall f \in R^n$, $f^T Lf = \frac{1}{2}\sum_{i,j=1}^n w_{ij}(f_i - f_j)^2$; b) L is symmetric and PSD; C) L has eigenvalues $0 = \lambda_1 \leq \cdots \leq \lambda_n$ with $(1, \cdots, 1)^T$ being the eigenvalue associated with 0.
  Proof: a) $f^T Lf = f^T Df - f^T Wf = \sum_i d_i f_i^2 - \sum_{i,j} w_{ij} f_i f_j = \frac{1}{2}(\sum_i \sum_j w_{ij} f_i^2 + \sum_i \sum_j w_{ij} f_j^2 - 2\sum_i \sum_j w_{ij} f_i f_j) = \frac{1}{2}\sum_{i,j} w_{ij}(f_i - f_j)^2$. b) Row/column sums of L are 0.
  Thm: The multiplicity of eigenvalue 0 of L equals the # of connected components of G, say $A_1, \cdots, A_k$. The corresponding eigenspace is spanned by $1_{A_1}, \cdots, 1_{A_k}$.
  Observation: The between-cluster dissimilarity is reflected by the smallest nonzero eigenvalue of L.
  Spectral Clsutering: 1) Fixed the m eigenvectors associated with the m smallest eigenvalues of L, denoted $Z_{n\times m}$; 2) Apply K-means to the rows of Z.
  Guarantees: 1) graph cut; 2) random walks; 3) matrix pertubation theory.

- Dimension reduction

  PCA: Idea: Find best rank-q of linear approximation to the data. Approximate x_i by $\mu + V_q\lambda_i, V_q \in M_{p\times q}(R), \lambda_i \in R^q$.
  Goal: Minimise $\sum_{i=1}^m ||x_i - \mu - V_q\lambda_i||^2$, V_q has orthogonal columns.
  One can show, given V_q, $\hat{\mu} = \bar{x}, \hat{\lambda}_i = V_q^T(x_i - \bar{x})$. Then find V_q minimising $\sum_{i=1}^n ||(x_i - \bar{x}) - V_q V_q^T(x_i - \bar{x})||^2$ or $||X - XV_q V_q^T||_F^2$. The solution is given by the SVD of X: $X = UDV^T$ and V_q is the first q columns of V. Columns of UD are called the principal components of X.

9 Graphical Models

- Gaussian graphical models

  $X = (X_1, \cdots, X_p) \sim N_p(\mu, \Sigma)$ where Σ is positive definite, undirected graph $G = (V, E), V = \{1, \cdots, p\}$ vertex set and E edge set $(i, j) \notin E$ iff $X_i \perp\!\!\!\perp X_j | X_{\{1,\cdots,p\}\setminus\{i,j\}}$ (conditional independence).

  Precision/concentration/inverse covariance matrix: $\Theta = \Sigma^{-1}$.

  Prop: $X_i \perp\!\!\!\perp X_j | X_{\{1,\cdots,p\}\setminus\{i,j\}}$ iff $\Theta_{ij} = 0$.

  Proof: By properties of multivariate normal, the distribution of $X_{(1)} = (X_i, X_j)$ given $X_{(2)} = X_{\{1,\cdots,p\}\setminus\{i,j\}}$ is $N_2(\mu_{1|2}, \Sigma_{1|2})$ where $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Thus $X_i \perp\!\!\!\perp X_j | X_{\{1,\cdots,p\}\setminus\{i,j\}}$ iff $\sigma_{ij} = 0$. On the other hand, by partioning $\Theta\Sigma = I$, $\Theta_{11}\Sigma_{11} + \Theta_{12}\Sigma_{21} = I, \Theta_{11}\Sigma_{12} + \Theta_{12}\Sigma_{22} = 0$. Then $\Theta_{11}\Sigma_{1|2} = \Theta_{11}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) = \Theta_{11}\Sigma_{11} - \Theta_{11}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = I - \Theta_{12}\Sigma_{21} + \Theta_{12}\Sigma_{21} = I$. So that $\Sigma_{1|2} = \Theta_{11}^{-1} = \frac{1}{\det(\Theta_{11})}\begin{pmatrix} \theta_{jj} & -\theta_{ij} \\ -\theta_{ji} & \theta_{ii} \end{pmatrix}$. This implies that $\sigma_{1|2,ij} = 0$ iff $\theta_{ij} = 0$.

- Precision matrix estimation

  Method 1: Neighborhood-based. From $X \sim N_p(0, \Sigma^{-1})$, we have $X_A | X_{A^c} \sim N(-\Theta_{AA}^{-1}\Theta_{AA^c}X_{A^c}, \Theta_{AA}^{-1})$, suggesting the linear model $X_A = B_A^T X_{A^c} + \eta_A$, where $B_A = -\Theta_{AA^c}\Theta_{AA}^{-1}$. When $A = \{i\}$, this reduces to $X_i = \beta_i^T X_{\{1,\cdots,p\}\setminus\{i\}} + \eta_i$ where $\beta_{ij} = -\frac{\theta_{ij}}{\theta_{ii}}$. "nodewise regression": $\mathrm{supp}(B) = \mathrm{supp}(\Theta)$.

  Pros: Borrow techniques from linear regression, fairly stable.

  Cons: Nontrivial to estimate magnitud of θ_{ij}. Not symmetric or positive definite.

  Method 2: Penalized likelihood (graphical Lasso): The Gaussian log-likelihood $l(\mu, \Sigma) = \frac{n}{2}\mathrm{logdet}(\Theta) - \frac{1}{2}\sum_{i=1}^n (x_i - \mu)\Theta(x_i - \mu)$. Substituing the MLDE $\bar{X}$ for μ, $l(\Theta) = \frac{n}{2}\mathrm{logdet}(\Theta) - \frac{n}{2}\mathrm{tr}(\Theta\frac{n}{2})$. Assume Θ is sparse, $\mathrm{minimise}_{\Theta \succ 0} - \mathrm{logdet}(\Theta) + \mathrm{tr}(\Theta\frac{n}{2}) + \lambda||\Theta||_1$(entrywise l_1-norm).

  Method 3: CLIME (Constrained l_1-minmimization): minimise $||\Theta||_1$ s.t.$||\hat{\Sigma}\Theta - I||_\infty \leq \lambda$. Equivalent to p linear programming probelms: minimise $||\theta_i||_1$ s.t.$||\hat{\Sigma}\theta_i - e_i||_\infty \leq \lambda \Rightarrow \widetilde{\Theta}$. Symmetrization: $\hat{\Theta} = (\hat{\theta}_{ij})$ with $\hat{\theta}_{ij} = \hat{\theta}_{ji} = \widetilde{\theta}_{ij}I(|\widetilde{\theta}_{ij}| \leq |\widetilde{\theta}_{ji}|) + \widetilde{\theta}_{ji}I(|\widetilde{\theta}_{ji}| > |\widetilde{\theta}_{ij}|)$.

  Nonasymptotic error bounds for CLIME: Sparsity class: $U_q(M, s_0(p)) = \{\Theta : \Theta \succ 0, ||\Theta||_{\mathscr{L}_1} \leq M, \max_i \sum_{j=1}^p |\theta_{ij}|^q \leq s_0(p), 0 \leq q < 1\}$.

  Lemma: Assume $\Theta_0 \in U_q(M, s_0(p))$. If $\lambda \geq ||\Theta_0||_{\mathscr{L}_1}||\hat{\Sigma} - \Sigma_0||_\infty$, then $||\hat{\Theta} - \Theta_0||_\infty \leq 4||\Theta_0||_{\mathscr{L}_1}\lambda, ||\hat{\Theta} - \Theta_0||_{\mathscr{L}_1} \leq Cs_0(p)\lambda^{1-q}$.

  Proof: Since $||\hat{\Sigma}\Theta_0 - I||_\infty = ||(\hat{\Sigma} - \Sigma_0)\Theta_0||_\infty \leq ||\hat{\Sigma} - \Sigma_0||_{\mathscr{L}_1}||\Theta_0||_\infty \leq \lambda, \Theta_0$ is a feasible solution. By the optimality of Θ, $\widetilde{\Theta}||_{\mathscr{L}_1} \leq ||\Theta_0||_{\mathscr{L}_1}$. Write that $||\widetilde{\Theta} - \Theta_0||_\infty = ||\Theta_0\Sigma_0(\widetilde{\Theta} - \Theta_0)||_\infty \leq ||\Theta_0||_{\mathscr{L}_1}||\Sigma_0(\widetilde{\Theta} - \Theta_0)||_\infty \leq ||\Theta_0||_{\mathscr{L}_1}\{||\hat{\Sigma}(\widetilde{\Theta} - \Theta_0)||_\infty + ||(\hat{\Sigma} - \Sigma_0)(||\widetilde{\Theta} - \Theta_0)||_\infty\} := ||\Theta_0||_{\mathscr{L}_1}(T_1 + T_2)$. Note that $T_1 \leq ||\hat{\Sigma}\widetilde{\Theta} - I||_\infty + ||\hat{\Sigma}\Theta_0 - I||_\infty \leq 2\lambda, T_2 \leq ||\widetilde{\Theta} - \Theta_1||_{\mathscr{L}_1}||\hat{\Sigma} - \Sigma_0||_\infty \leq 2||\Theta_0||_{\mathscr{L}_1}||\hat{\Sigma} - \Sigma_0||_\infty \leq 2\lambda$. Thus, $||\widetilde{\Theta} - \Theta_0||_\infty \leq 4||\Theta_0||_{\mathscr{L}_1}\lambda$ which implies $||\hat{\Theta} - \Theta_0||_\infty \leq 4||\Theta_0||_{\mathscr{L}_1}\lambda$.

  Let $t_n = ||\hat{\Theta} - \Theta_0||_\infty, \delta_i = \hat{\theta}_i - \theta_i^0 = \delta_i^{(1)} + \delta_i^{(2)}$, where $\delta_{ij}^{(1)} = \hat{\theta}_{ij}I(|\hat{\theta}_{ij}| \geq 2t_n) - \theta_{ij}^0, \delta_{ij}^{(2)} = \hat{\theta}_{ij}I(|\hat{\theta}_{ij}| \leq 2t_n)$. Then $||\theta_i^0||_1 \geq ||\hat{\theta}_i||_1 = ||\theta_i^0 + \delta_i^{(1)}|| + ||\delta_i^{(2)}||_1 \geq ||\theta_i^0||_1 - ||\delta_i^{(1)}||_1 + ||\delta_i^{(2)}||_1$, so that $||\delta_i^{(2)}||_1 \leq ||\delta_i^{(1)}||_1$ and hence $||\delta_i||_1 \leq 2||\delta_i^{(1)}||_1$. By the sparsity assumption, $||\delta_i^{(1)}|| = \sum_{j=1}^p |\hat{\theta}_{ij}I(|\hat{\theta}_{ij}| \geq 2t_n) - \delta_{ij}^0| \leq \sum_{j=1}^p |\theta_{ij}^0|I(|\theta_{ij}^0| < 2t_n) + \sum_{j=1}^p |\hat{\theta}_{ij}I(|\hat{\theta}_{ij}| \geq 2t_n) - \hat{\theta}_{ij}^0 I(|\theta_{ij}^0| \geq 2t_n) \leq (2t_n)^{1-q}\sum_{j=1}^p |\theta_{ij}^0|^q + \sum_{j=1}^p |\hat{\theta}_{ij} - \theta_{ij}^0|I(|\hat{\theta}_{ij}| \geq 2t_n) + \sum_{j=1}^p |\theta_{ij}^0||I(|\hat{\theta}_{ij}| \geq 2t_n) - I(|\theta_{ij}^0| \geq 2t_n)| \leq (2t_n)^{1-q}s_0(p) + t_n\sum_{j=1}^p I(|\theta_{ij}^0| \geq t_n) + \sum_{j=1}^p |\theta_{ij}^0|I(|\theta_{ij}^0 - 2t_n| \leq |\hat{\theta}_{ij} - \theta_{ij}^0|) \leq (2t_n)^{1-q}s_0(p) +$

$t_n^{1-q}\sum_{j=1}^p|\theta_{ij}^0|^q+\sum_{j=1}^p|\theta_{ij}^0|I(|\theta_{ij}^0|\leq 3t_n)=(2t_n)^{1-q}s_0(p)+t_n^{1-q}s_0(p)+(3t_n)^{1-q}s_0(p)=(1+2^{1-q}+3^{1-q})t_n^{1-q}s_0(p)$. Combine the above to conclude: $||\hat{\Theta}-\Theta_0||_{\mathscr{L}_1}=\max_i||\delta_i||_1\leq 2(1+2^{1-q}+3^{1-q})(4||\Theta_0||_{\mathscr{L}_1})^{1-q}s_0(p)\leq Cs_0(p)\lambda^{1-q}$.

Thm: $||\hat{\Theta}-\Theta_0||_{\mathscr{L}_1}\leq C_1M^{1-q}s_0(p)(\frac{\log p}{n})^{\frac{1-q}{2}}$ with high probability.

Proof: One can show that $||\hat{\Theta}-\Theta_0||_\infty\leq C_2\sqrt{\frac{\log p}{n}}$ with high probability. Take $\lambda=C_2M\sqrt{\frac{\log p}{n}}$.

- Directed acyclic graphs

Def: $G=(V,E),V=\{1,2,\cdots,p\},E\subset V\times V,i\to j:(i,j)\in E,i\leftrightarrow j:(i,j)\&(j,i)\in E$. A DAG is a graph whose all edges are directed and that contains no cycles.

Parents of node $j:\text{pa}(j)=\{i\in V:i\to j\}$; adjacency set: $\text{adj}(j)=\{i\in V:j\to i \text{ or } i\to j\}$. A distribution f factorizes with regard to G iff $f(x)=\prod_{v\in V}f(x_v|x_{\text{pa}(v)})$.

Undirected graph $x-y-z$, orientation: $x\to y\to z,z\to y\to x,z\leftarrow y\to x,z\to y\leftarrow x$. $(0)=f(x,y,z)=f(x)f(y|x)f(z|y),(1)=f(x)f(y|x)f(z|y),(2)=f(z)f(y|z)f(x|y)=f(y)f(z|y)f(x|y)=f(x)f(y|x)f(z|y),(3)=f(y)f(z|y)f(x|y)=f(x)f(z|y)f(y|x),(4)=f(x)f(z)f(y|x,z)$.

Thm: Two DAGs are Markov equivalent iff they have the sae skeleton and V-structures. Skeleton: undirected graph replacing all directed edges with undirected ones. V-structures: unshielded collider ($x\to z\leftarrow y$ but $x-\times-y$).

PDAG (Partially DAG): All V-structures are oriented.

Completed PDAG: Besides V-structures, oriented as much as possible.

Def: d-separation: A path is d-separated by s set of notes Z iff it contains either: 1) a chain $i\to m\to j$ or fork $i\leftarrow m\to j$ s.t. $m\in Z$; 2) a V-structure $i\to m\leftarrow j$ s.t. $m\notin Z$ and no descendant of m belongs to Z. Z d-separates x from y iff Z d-separates every path from x to y.

Faithfulness: A prob dist P on R^p is faithful to G iff $\forall i,j\in V,i\neq j,\ \&\,S\subset V,X_i\perp\!\!\!\perp X_j|X_S\Leftrightarrow i\,\&\,j$ are d-separated by S.

PC algorithm: for estimating a DAG from the data.

Stage 1: Find the skeleton. Check if $x_i\perp\!\!\!\perp x_j|S,\forall S\subset V\backslash\{i,j\}$.

```
Part 1: $l=0,G=$ complete graph.
Choose $(i,j)\in E(G)$ s.t. $|\text{adj}(G,i)\backslash\{i\}|\geq l$.
—— Choose $K\subset \text{adj}(G,i)\backslash\{j\}$ with $|K|=l$.
—— —— Test if $x_i\perp\!\!\!\perp x_j|x_K$. If yes, then a) delete $(i,j)$ from $E(G)$; b) save $K$ in $S(i,j)$ and $S(j,i)$.
—— until no such $K$ is left.
—— $l\leftarrow l+1$.
until no such $(i,j)$ is left.
```

Prop: $\forall i,j\in V,K\subset V\backslash\{i,j\},h\in K,\rho_{i,j|K}=\frac{\rho_{i,j|K\backslash\{h\}}-\rho_{i,h|K\backslash\{h\}}\rho_{j,h|K\backslash\{h\}}}{\sqrt{(1-\rho^2_{i,h|K\backslash\{h\}})(1-\rho^2_{j,h|K\backslash\{h\}})}}$. Z-transform: $Z(i,j|K)=\frac{1}{2}\log(\frac{1+\hat{\rho}_{i,j|K}}{1-\hat{\rho}_{i,j|K}})$, reject $\mathcal{H}_0:\rho_{i,j|K}=0$ if $\sqrt{n-|K|-3}Z(i,j|K)>\Phi^{-1}(1-\frac{\alpha}{2})$.

Stage 2: Extend the skeleton to a CPDAG. $\forall$ nonadj i and j with common neighbor k,

R 0): If $k\neq S(i,j)$, then $i\leftarrow k\to j$. Orient as many edges as possible in the PDAG.

R 1): Away from V-structures. $j\to k$ whenever $i\to j-k,i-\times-k$.

R 2): Away from cycles: $i\to j$ whenever $i\to k\to j,i-j$.

R 3): Double triangle: $i\to j$ whenever $i-k\to j,i-l\to j,i-j$.

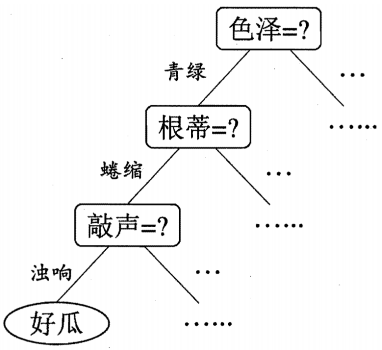
Complexity: worse case: $O(n(p^q\vee p^2))$ where $|\text{adj}(i)|\leq q$.

10 Random Forests

- Decision trees

  Idea: Partition feature space sequentially, each time by a single variable.

  Parameters: splittig variables/points, tree topology, decision on each leaf.



- Regression trees

  Given partitions $\{R_1, \cdots, R_M\}$, $f(x) = \sum_{m=1}^{M} c_m I(x \in R_m)$.

  $\hat{c}_m = \text{ave}(y_i | x_i \in R_m)$.

  Greedy strategy: Define $R_1(j,s) = \{x : x_j \leq s\}, R_2(j,s) = \{x : x_j > s\}$.

  $(\hat{j}, \hat{s}) = \text{argmin}_{(j,s)}\{\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2\}$.

  $\hat{c}_1 = \text{ave}(y_i | x_i \in R_1(j,s)), \hat{c}_2 = \text{ave}(y_i | x_i \in R_2(j,s))$.

  Regularization: pre-pruning, cost-complexity pruning.

  Cost function: $Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2$ where $N_m = \#\{x_i \in R_m\}$.

  Model complexity: $|T| = \#$ regions.

  Criterion: $C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T|$.

- Classification trees

  Decision on each R_m: majority vote. $k(m) = \text{argmax}_k \hat{p}_{mk}$ where $p_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$.

  Measure of node impurity: misclassification error: $\frac{1}{N_m} \sum_{x_i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{m,k(m)}$.

  Gini-Simpson index: $\sum_{(j,k):j \neq k} \hat{p}_{m_j} \hat{p}_{m_k} = 1 - \sum_k \hat{p}_{m_k}^2$.

  Shannon index/cross-entropy/deviance: $-\sum_k \hat{p}_{m_k} \text{log} \hat{p}_{m_k}$.

- Ensemble learning

  Two types:

  1) strongly dependant, sequentially trained (boosting);

  2) weekly dependant, parallelly trained (bagging, RF).

  Bagging (bootstrap aggregation): Traning data $(x_i, y_i), i = 1, \cdots, n$, generate B bootstrap samples.

  1) regression: $\hat{f}_{\text{bag}}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b^*(x)$;

  2) classification: $\mathcal{H}(x) = \text{argmax}_k \sum_{b=1}^{B} I(h_b^*(x) \neq k)$.

- Random forests

  A varient of bagging on decorrelated trees as base learners, with additional randomization step: choose randomly $m \leq p$ features before each splitting.

11 Reinforcement Learning

- General scenario of RL

  Agent → (action) → Environment → (state, reward) → Agent. Tradeoff between exploration and exploitation. Goal: Determine the optimal policy (course of actions) to maximize its reward.

  Settings: Environment model known – Planning; unknown – Learning.

  Basic assumption: Markov decision process (MDP).

  State $s \in S$, initial state s_0, action $a \in A$, transition probabilities $P(s'|s,a)$, reward probabilities $P(r|s,a)$. $s_t \to a_t/r_t \to s_{t+1} \to a_{t+1}/r_{t+1} \to s_{t+2} \to \cdots$.

  Def: Policy $\pi : S \to D(A)$ (distribution on A). Deterministic if $\pi(s)(a) = 1$ for some a. Stationary π, not dependant on t; non-stationary π_t.

  Return: finite horizon $T < \infty$: $\sum_{t=0}^{T} r(s_t, \pi(s_t))$, $T = \infty$: $\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t))$.

  Policy value (expected return): $T < \infty : V_\pi(s) = E_{a_t \sim \pi(s_t)}(\sum_{t=0}^{T} r(s_t, a_t)|s_0 = s)$;

  $T = \infty : E_{a_t \sim \pi(s_t)}(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|s_0 = s)$ $(0 < \gamma < 1)$.

  Optimal policies: π^* is optimal if $\forall \pi \& s \in S, V_{\pi^*}(s) \geq V_\pi(s)$.

  Def: State-action value function: $Q_\pi(s,a) = E(r(s,a) + \gamma V_\pi(s_1)|s_0 = s, a_0 = a)$ (first take action a + then follow policy π).

  Thm: (Policy improvement) $\forall$ policy π, π', $[\forall s \in S, E_{a \sim \pi'(s)} Q_\pi(s,a) \geq E_{a \sim \pi(s)} Q_\pi(s,a)] \Rightarrow [\forall s \in S, V_{\pi'}(s) \geq V_\pi(s)]$.

  Proof: LHS $V_\pi(s) = E_{a \sim \pi(s)} Q_\pi(s,a) \leq E_{a \sim \pi'(s)} Q_\pi(s,a) = E_{a \sim \pi'(s)}(r(s,a) + \gamma V_\pi(s_1)|s_0 = s) = E_{a \sim \pi'(s)}(r(s,a) + \gamma E_{a_1 \sim \pi'(s_1)} Q_\pi(s,a)|s_0 = s) \leq \cdots \leq E_{a_t \sim \pi'(s_t)}(\sum_{t=0}^{T} \gamma^t E r(s_t, a_t) + \gamma^{T+1} V_\pi(s_{T+1})|s_0 = s)$. Let $T \to \infty$.

  Corollary (Bellman's optimality condition): π is optimal iff $\forall V(s,a)$ with $\pi(s)(a) > 0, a \in \mathrm{argmax}_{a' \in A} Q_\pi(s,a')$.

  Thm: Any finite MDP admits an optimal deterministic policy.

  Proof: Consider the deterministic policy π^* maximizing $\sum_{s \in S} V_\pi(s)$, π^* exists since finitely many. If π^* were not optimal, then it could be imporved by some s with $\pi(s) \notin \mathrm{argmax}_{a' \in A} Q_\pi(s,a')$.

  Bellman's equations: 1) For the optimal policy value $V^*(s) = Q^*(s, \pi^*(s))$, $V^*(s) = \max_{a \in A}(E r(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^*(s'))$; 2) For general $V_\pi(s)$, $V_\pi(s) = E_{a_1 \sim \pi(s)} r(s,a_1) + \gamma \sum_{s' \in S} P(s'|s,\pi(s)) V_\pi(s')$. In matrix form, $V = R + \gamma P V$ or $V = (I - \gamma P)^{-1} R$.

  Why $I - \gamma P$ invertible? P stachastic matrix, $||P||_\infty = 1$, and hence $||\gamma P||_2 \leq ||\gamma P||_\infty = \gamma < 1$.

- Planning algorithms

  Let $\Phi(V) = \max_\pi(R_\pi + \gamma P_\pi V)$.

  Value iteration algorithm:

  $V = V_0$. While $||V - \Phi(V)|| \geq \frac{1-\gamma}{\gamma} \epsilon$ do $V \leftarrow \Phi(V)$

  Thm: Converges $\forall V_0$.

  Q-learning:

  Sample a new state s'; update policy values by $Q(s,a) \leftarrow (1-\alpha) Q(s,a) + \alpha(r(s,a) + \gamma \max_{a' \in A} Q(s',a'))$ (stochastic approximation).

  Thm: Converges for finite MDPs whenever $\sum_{t=0}^{\infty} \alpha_t(s,a) = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2(s,a) < \infty$.