CHAPTER 6: BEYOND PARAMETRIC MODELS AND BEYOND ESTIMATION

INTRODUCTION TO NONPARAMETRIC/SEMIPARAMETRIC MODELS

Nonparametric/Semiparametric Estimation

- Parametric models uses only a finite number of parameters to describe data distribution.
- Model parameters are convenient for interpretation.
- However, they are not sufficiently accurate to describe complex data generation.
- Model misspecification can lead to severe bias or incorrect inference.
- More flexible models include nonparametric and semiparametric models.

Nonparametric density estimation

- One fundamental problem in statistical inference is density estimation.
- Parametric models can be normal distribution, $t-$distribution and etc.
- Nonparametric model requires no assumption on the form of density functions.
- Assume i.i.d. observations $X_1, ..., X_n$ from a distribution with density $f(x)$.
- The goal is to estimate $f(x)$ without any assumptions.

Local approaches

- The idea is to estimate the density at any fixed x locally.
- Essentially, only observations close to x will contribute to estimation.
- Weights will be introduced to determine the locality of the observations.
- $$\widehat{f}(x) = n^{-1} \sum_{i=1}^{n} w_{ni}(x),$$

  where
  $$w_{ni}(x) = a_n^{-1} K\left(\frac{X_i - x}{a_n}\right)$$
  and $K(x) \geq 0$ satisfying $\int K(x)dx = 1$.
- a_n is called the bandwidth.

Justification

- Show $E[\widehat{f}(x)] \to f(x)$ when $a_n \to 0$.
- Bias analysis

$$E[\widehat{f}(x)] - f(x) = \int_y K(y) f(x + a_n y) dy - f(x).$$

- Variance analysis

$$Var[\widehat{f}(x)^2] = (na_n)^{-1} \left[\int K(y)^2 f(x + a_n y) dy \right.$$

$$\left. - a_n \left(f(x) + \mathrm{Bias} \right)^2 \right].$$

Some conclusions

- If $K(x) = 0.5I(|x| \leq 1)$,

$$\widehat{f}(x) = (2a_n)^{-1}\left\{\widehat{F}(x + a_n) - \widehat{F}(x - a_n)\right\}.$$

- Bias$=f(x)a_n + O(a_n)$ and
  Variance$=(na_n)^{-1}f(x)\int K(y)^2dy + o((na_n)^{-1})$.
- If $K(x)$ is symmetric (Gaussian kernel or Epanechnikov kernel), then
  Bias$=a_n^2 f''(x)\int K(y)y^2dy/2 + o(a_n^2)$ and Variance remains the same.
- The choice of the kernel depends on how much smoothness is known about the density function.

Asymptotic normality

- $$\frac{\widehat{f}(x) - E[\widehat{f}(x)]}{\sqrt{Var(\widehat{f}(x))}} \rightarrow_d N(0,1).$$
- The proof assumes $na_n^3 \rightarrow 0$ and uses Liaponov CLT.
- For a symmetric kernel, the optimal bandwidth is

$$a_n^{optimal} = \left[\frac{4f(x)\int K(y)^2 dy}{(f''(x)\int K(y)y^2 dy)^2}\right]^{1/5} n^{-1/5}.$$

Global approaches

- It views $f(x)$ as a function parameter for estimation so estimates $f(x)$ via one global optimization instead of estimation at each x.
- It is computationally efficient.
- The disadvantage is that it may miss some local features of $f(x)$.

Empirical distribution function

- Instead of estimating density function, we estimate its distribution function $F(x)$.
- We consider maximizing the log-likelihood function

$$\sum_{i=1}^{n} \log f(X_i)$$

but replace $f(X_i)$ by

$$F\{X_i\} = F(X_i) - F(X_i-).$$

Asymptotic properties

- $\widehat{F}(x)$ converges to $F(x)$ almost surely.
- $$\sup_x |\widehat{F}(x) - F(x)| \to 0$$
  almost surely.
- $\sqrt{n}(\widehat{F}(x) - F(x))$ converges in distribution to a Brownian bridge process.
- The previous kernel density estimator can be viewed as a smoothing operation on $\widehat{F}$:

$$\widehat{f}(x) = \int a_n^{-1} K((y-x)/a_n) d\widehat{F}(y).$$

Sieve Estimation

- We approximate $f(x)$ via a sequence of functions generated from basis functions:

$$\log f(x) \approx \sum_{k=1}^{K_n} \beta_k B_k(x).$$

- Choices of basis functions: piecewise constant, piecewise linear, piecewise polynomials (splines), wavelets, trigonometric functions ...
- We then maximize the likelihood function subject to constraint $\int f(x)dx = 1$.
- When the number of basis function goes to infinity, the bias due to approximation will vanish.
- However, more basis functions will result in increasing variability.
- Asymptotic bias/variance analysis (also normality) is more complicated than and is not as obvious as local approaches.

Penalization approach

- The essential idea is to construct "Objective function" plus "Regularization" (penalty).
- The objective function is an empirical version of a population quantity which the true density function minimize.
- The regularization is a penalty function to penalize those estimators with high variability or irregularity.
- The common estimation is

$$\min -\sum_{i=1}^{n} \log f(X_i) + \lambda_n P(f), \quad \int f(x) = 1,$$

$$P(f) = \int |f''(x)|^2 dx.$$

- λ_n is the penalty parameter (tuning parameter) to govern the regularity of the estimator.
- Bias and variance trade-off is reflected in λ_n.

Nonparametric Regression

- The goal is to estimate the conditional mean of Y given X, $m(x) = E[Y|X = x]$.
- The data are $(Y_1, X_1), ..., (Y_n, X_n)$.
- Parametric models: linear model, generalized linear models
- Parameter models are easy for interpretation but can be seriously misspecified.

Nonparametric approaches

- Local approach (kernel estimation)

$$\frac{\sum_{i=1}^{n} Y_i K((X_i - x)/a_n)}{\sum_{i=1}^{n} K((X_i - x)/a_n)}.$$

- Local likelihood approach

$$\min \sum_{i=1}^{n} (Y_i - m(x))^2 K((X_i - x)/a_n).$$

- Local polynomials

Global approaches

- Sieve estimation

$$\min \sum_{i=1}^{n} (Y_i - \sum_{k=1}^{K_n} \beta_k B_k(X_i))^2.$$

- Penalization estimation

$$\min \sum_{i=1}^{n} (Y_i - m(X_i))^2 + \lambda_n P(m).$$

Semiparametric Estimation

- It aims to incorporate advantages from both parametric and nonparametric models.
- Recall: parametric models are easy for interpretation and estimation is precise with a finite number of parameters; nonparametric models are robust with minimal assumptions.
- Semiparametric models describe data distributions using both parametric components (θ) and nonparametric components (η).
- θ is finite dimensional and consists of parameters of interest (for convenience of practical use): treatment effects, risk ratios ...
- η is nonparametric and included to complement θ for describing data distribution. It is not the primary interest so called nuisance parameters.

Inferential advantage and challenges

- Most often, the parameter θ can be estimated as accurately as from a parametric models (parametric convergence rate).
- The nuisance parameter, η, has minimal assumption so the inference is robust to the structure in η.
- Estimation/inference is challenging due to the mixing nature of the parameters.
- Usually, we have to treat η as some parameter from a metric space for inference. Some math from function analysis is quite involved.

Examples

- Right censored data
- Current status data
- Smoking prevention project
- Medical cost

Estimation approaches

- Direct plug-in estimation of nuisance parameters
- Estimating equations
- IPWE for missing data
- NPMLE approach
- Profile likelihood estimation
- Sieve estimation
- Penalization estimation

INTRODUCTION TO STATISTICAL LEARNING

Statistical Learning

- What is statistical learning?
  - machine learning, data mining
  - supervised vs unsupervised

- How different from traditional inference?
  - different objectives
  - different statistical procedures
  - supervised learning $< - - - >$ regression
  - unsupervised learning $< -- >$ density estimation

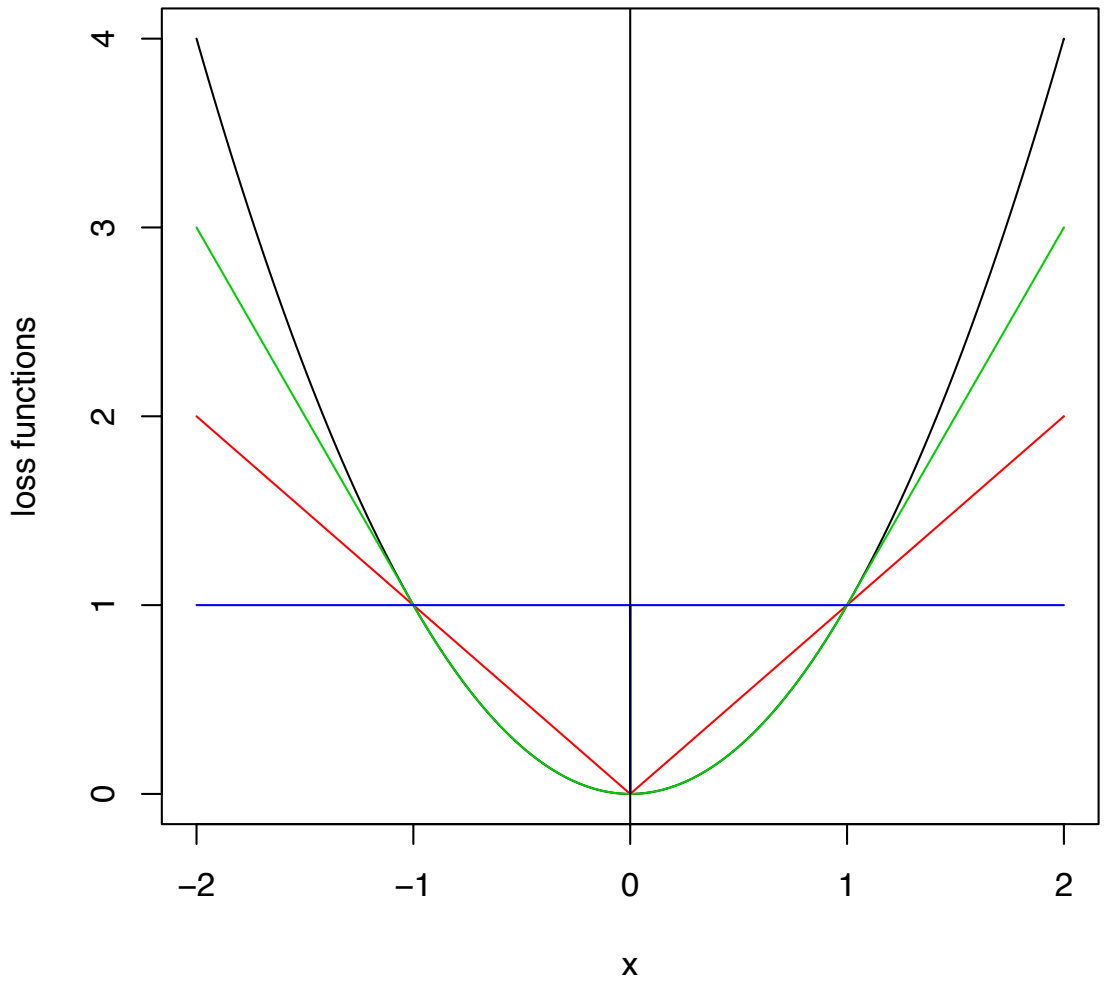
Set-up in decision theory

- X: feature variables
- Y: outcome variable (continuous, categorical, ordinal)
- (X, Y) follows some distribution
- goal: determine $f : X \to Y$ to minimize some loss

$$E[L(Y, f(X))].$$

Loss function $L(y, x)$

- squared loss: $L(y, x) = (y - x)^2$
- absolute deviation loss: $L(y, x) = |y - x|$
- Huber loss: $L(y, x) = (y - x)^2 I(|y - x| < \delta) + (2\delta|y - x| - \delta^2) I(|y - x| \geq \delta)$
- zero-one loss: $L(y, x) = I(y \neq x)$
- preference loss: $L(y_1, y_2, x_1, x_2) = 1 - I(y_1 < y_2, x_1 < x_2)$


loss functions
0
1
2
3
4
−2
−1
0
1
2
x

Optimal $f(x)$

- squared loss: $f(X) = E[Y|X]$
- absolute deviation loss: $f(X) = med(Y|X)$
- Huber loss: ???
- zero-one loss: $f(X) = \text{argmax}_k P(Y = k|X)$
- preference loss: ???
- not all loss functions have explicit solutions

Estimate $f(x)$

- Empirical data

$$(X_i, Y_i), \quad i = 1, ..., n$$

- Direct learning: estimate f directly via parametric, semi-parametric, or nonparametric methods
- Indirect learning: estimate f by minimizing (empirical risk)

$$\sum_{i=1}^{n} L(Y_i, f(X_i))$$

Candidate set for $f(x)$

- too small: underfit data
- too large: overfit data
- even more important with high-dimensional X

Why high-dimensionality is an issue?

- data are sparse
- local approximation is infeasible
- increasing bias and variability with dimensionality
- curse of dimensionality

Common considerations for $f(x)$

- linear functions or local linear functions
- linear combination of basis function: polynomials, splines, wavelets
- let data choose f by penalizing f from roughness

Parametric learning

- It is one of direct learning methods.
- Estimate $f(x)$ using parametric models.
- Linear models are often used.

Linear regression model

- Target squared loss or zero-one loss.
- Assume $f(X) = E[Y|X] = X^T\beta$.
- The least squared estimation

$$\hat{f}(x) = x^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y.$$

Shrinkage methods

- Gain variability reduction by sacrificing prediction accuracy.
- Help to determine important features (variable selection) if any.
- Include subset selection, ridge regression, LASSO and et.

Subset selection

- Search for the best subset of size k in terms of RSS.
- Use leaps and bounds procedure.
- Computationally intensive with large dimension.
- The best choice of size k is based on Mallow's CP.

Ridge regression

- Minimize

$$\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \lambda\sum_{j=1}^{p}\beta_j^2.$$

- Equivalently, minimize

$$\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2, \quad \text{subject to } \sum_{j=1}^{p}\beta_j^2 \leq s.$$

- The solution

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

- Has Bayesian interpretation.

LASSO

- Minimize

$$\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|.$$

- Equivalently, minimize

$$\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2, \quad \text{subject to } \sum_{j=1}^{p}|\beta_j| \leq s.$$

- This is a convex optimization.
- Suppose $\mathbf{X}$ to have independent columns:

$$\hat{\beta}_j = \text{sign}(\hat{\beta}^{lse})(|\hat{\beta}^{lse}| - \lambda/2)^+.$$

- Nonlinear shrinkage property.

Summary

- Subset selection is L_0-penalty shrinkage but computationally intensive.
- Ridge regression is L_2-penalty shrinkage and shrinks all coefficients the same way.
- LASSO is L_1-penalty shrinkage and it is a nonlinear shrinkage.

Other shrinkage methods

- L_q-penalty with $q \in [1, 2]$:

$$\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q.$$

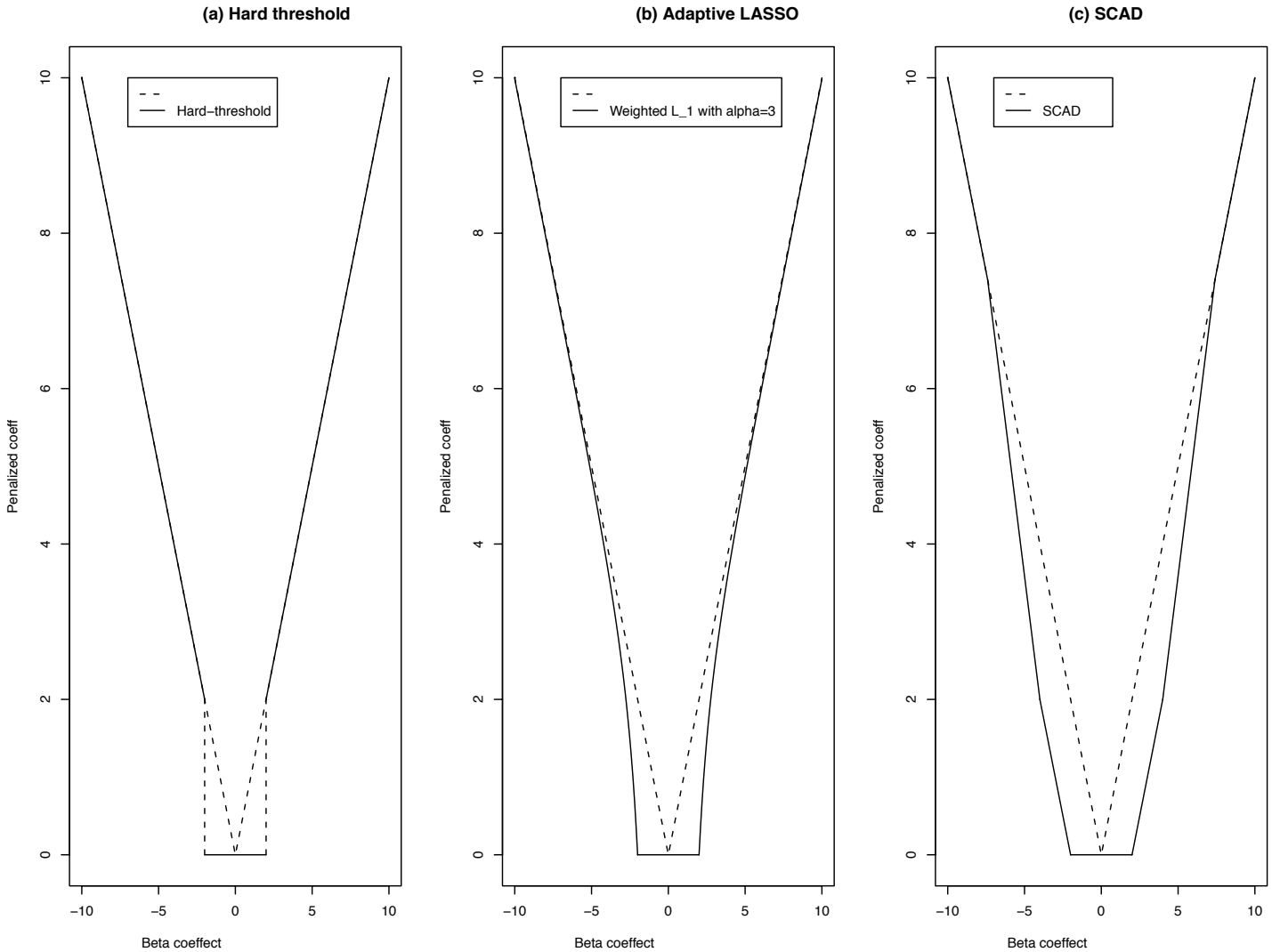
- Weighted LASSO (aLASSO):

$$\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \lambda \sum_{j=1}^{p} w_j|\beta_j|$$

  where $w_j = |\hat{\beta}^{lse}|^{-q}$.

- SCAD penalty $\sum_{j=1}^{p} J_\lambda(|\beta_j|)$:

$$J'_\lambda(x) = \lambda \left\{ I(x \leq \lambda) + \frac{(a\lambda - x)_+}{(a-1)\lambda} I(x > \lambda) \right\}.$$


(a) Hard threshold
Hard−threshold
Penalized coeff
Beta coeffect
(b) Adaptive LASSO
Weighted L_1 with alpha=3
Penalized coeff
Beta coeffect
(c) SCAD
SCAD
Penalized coeff
Beta coeffect

Compare different penalties

- All penalties have shrinkage properties.
- Some penalties give an oracle property as if the true zeros are known (aLASSO, SCAD).
- But aLASSO needs a consistent initial estimate (not suitable for high-dimensional).
- SCAD generally needs large sample size and may suffer computational difficulty (due to its non-convexity).

Logistic discriminant analysis

- It is often used when Y is dichotomous or categorical.
- Assume

$$P(Y = k|X = \frac{\exp\{\beta_{k0} + X^T\beta_k\}}{1 + \sum_{l=1}^{K} \exp\{\beta_{l0} + X^T\beta_l\}}.$$

- Then

$$\hat{f}(x) = \text{argmax}_k \left\{\hat{\beta}_{k0} + X^T\hat{\beta}_k\right\}.$$

Discriminant analysis

- Assume that X given $Y = k$ follows a normal distribution with mean μ_k and covariance Σ_k.
- For $K = 2$, the decision rule (quadratic discriminant analysis) is based on the sign of

$$\log \frac{\pi_2}{\pi_1} - \frac{1}{2}(x-\hat{\mu}_2)^T \hat{\Sigma}_2^{-1}(x-\hat{\mu}_2) + \frac{1}{2}(x-\hat{\mu}_1)^T \hat{\Sigma}_1^{-1}(x-\hat{\mu}_1).$$

- If assume $\Sigma_1 = \Sigma_2$, this results in linear discriminant anaysis.

Generalization

- In parametric methods, features X can be replaced by some basis functions so we have nonlinear discriminant boundary.
- Efficient estimation for $f(x)$ is possible due to parametric nature.

What is semi-nonparametric?

- It is neither parametric nor nonparametric.
- But it is also difference from usual semiparametric models.
- It includes neural networks, projection pursuit, GAM and MARS.

Neural network

- It is an artificially structured model.
- Assume one or more hidden layers between input X and output Y.
- Simple models between one layer variables and its upper layer.
- Forward- and backward-propagation algorithms are used for calculation.

Generalized additive models

- $f(x)$ is assumed to take form

$$\sum_{k=1}^{p} f_k(X_{(k)}.$$

- More flexible than parametric models
- But assume no interactions among X's.
- Backfitting is used for estimation, where each step is a univariate nonparametric estimation.
- It applies for continuous and categorical outcome variable.

Projection pursuit

- $f(x)$ takes form

$$\sum_{k=1}^{m} g_k(\beta_k^T X).$$

- More general than GAM.
- Include single index model as special cases and allow X's interactions.
- Recursively estimate each single-index component.
- A local linear approximation and backfitting are used for each step.

Direct learning: nonparametric approaches

- No structural assumption for $f(x)$.
- They strongly relate to nonparametric regression in traditional statistical estimation.
- Include k-NN, kernel methods, sieve methods, tree methods and MARS.

Nearest neighbor methods

- It is a prototype method.
- The estimation is the majority of outcomes in k-neighborhood.
- Distance is an important issue in defining neighborhood.
- Classification boundary is usually irregular.

Kernel methods

- It is one of the most popular methods in nonparametric estimation.
- Estimation is based on a locally weighed average, where weights are given by some kernel function.
- One important issue is the choice of the bandwidth (bias and variance tradoff).
- It is equivalent to a local constant estimation.
- Generalized to local linear and local polynomials.

Sieve methods

- It is a global approximation to $f(x)$.
- The idea is simple: approximate $f(x)$ by a series of basis functions.
- The choices of basis functions: polynomials, trigonometric functions, regression splines, B-splines, wavelets.
- The choices of the number of basis functions is important.
- Adapt to specific applications.

Tree methods

- Regression tree for continuous Y and classification tree for categorical Y.
- It is a sequentially and recursively partition of X's space.
- Each partition is done for one X's component and the partition is usually binary.
- The way of choosing which X and where for partition relies on some specific criteria.
- The tree can grow to the full length but needs pruning to avoid overfitting.
- Tree size is often chosen as a way to prune the tree.
- A generalization is called random forest: a bootstrapped way of growing tree to avoid over-dependence on one single tree.

Multivariate adaptive regression splines (MARS)

- Some combination of sieve methods and tree methods.
- The basis functions take form $(X_{(k)} - t)_+$ or $(t - X_{(k)})_+$ along with their interactions.
- Like the tree, it is a sequential fitting method.
- A backward deletion procedure is applied to avoid overfitting.

Which methods should we choose?

- It depends on specific data and applications.
- Kernel and spline methods are useful for smooth signal and possess nice theoretical properties.
- Wavelets are useful for discontinuous signal (denoise imaging).
- Tree methods and MARS have computational advantages and decision rules are simple but both lack nice theoretical properties.
- Tree methods are applicable to high-dimensional X.

Indirect learning

- It doesn't estimate $f(x)$ directly, most likely due to in-explicit $f(x)$.
- It estimates the decision rule through minimizing empirical risks.
- It includes SVM and regularized minimization.

Support vector machine

- Assume $Y \in \{-1, 1\}$.
- The goal is to find a hyperplane $\beta_0 + X^T\beta$ which can separate Y's maximally.
- That is, we wish

$$Y_i(\beta_0 + X_i^T\beta) > 0$$

for all $i = 1, ..., n$.

Perfect separation

- Consider an ideal situation where Y's can be perfectly separated.
- A maximal separation can be determined as that we want the minimum distance from each point to the separating plane as large as possible.
- It is equivalent to

$$\max_{\|\beta\|=1} C, \quad \text{subject to } Y_i(\beta_0 + X_i^T \beta) \geq C, i = 1, ..., n.$$

- The dual problem is

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j Y_i Y_j X_i^T X_j, \quad \alpha_i \geq 0.$$

Imperfect separation

- In real data, there is usually no hyperplane separating perfectly (if there is, it is by chance).
- We should allow some violations by introducing slack variables $\xi_i \geq 0$:

$$\max_{\|\beta\|=1} C, \text{ subject to } Y_i(\beta_0 + X_i^T \beta) \geq C(1-\xi_i) i = 1, ..., n.$$

- $\sum_i \xi_i$ describes the total degree of violation should be controlled (like type I error in hypothesis test):

$$\sum_i \xi_i \leq \text{ a given constant.}$$

Imperfect separation

- The dual problem is

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j Y_i Y_j X_i^T X_j,$$

$$0 \leq \alpha_i \leq \gamma, \quad \sum_{i=1}^{n} \alpha_i Y_i = 0.$$

- It is a convex optimization problem.
- It turns out $\hat{\beta} = \sum_{\hat{\alpha}_i > 0} \hat{\alpha}_i Y_i X_i$ so $\hat{\beta}$ is determined by the points within or on the boundary of a band around the hyperplane.
- These points are called support vectors.

SVM allowing nonlinear boundary

- Linear boundary may not be practical.
- To allow nonlinear boundary, assume

$$f(x) = (h_1(x), ..., h_m(x))\beta + \beta_0.$$

- The dual problem becomes

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j Y_i Y_j K(X_i, X_j),$$

$$0 \leq \alpha_i \leq \gamma, \quad \sum_{i=1}^{n} \alpha_i Y_i = 0.$$

- Here, $K(x, x') = (h_1(x), ..., h_m(x)(h_1(x'), ..., h_m(x'))^T$.
- Moreover,

$$\hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_i Y_i K(x, X_i) + \hat{\beta}_0.$$

- Thus, we only need to specific the kernel function $K(x, y)$.

Equivalent form of SVM

- SVM learning is equivalent to minimizing

$$\sum_{i=1}^{n} \{1 - Y_i f(X_i)\}_+ + \lambda \|\beta\|^2/2.$$

- Thus, it is a regularized empirical risk minimization.
- This formation is useful for justifying SVM's theoretical property.
- Other loss functions are possible.

Regularized estimation

- It is typically formed as

$$\min_{f \in \mathcal{H}} \left[\sum_{i=1}^{n} L(Y_i, f(X_i)) + \lambda J(f) \right].$$

- $J(f)$ penalizes those band f in $\mathcal{H}$.
- For example, $J(f) = (f''(x))^2 dx$ gives cubic spline approximation.
- More general, choose $\mathcal{H}$ to be a reproducing kernel Hilbert space and $J(f) = \|f\|_{\mathcal{H}_k}$.
- Then the problem becomes minimizing

$$\sum_{i=1}^{n} L(Y_i, \sum_{j=1}^{n} \alpha_j K(X_j, X_i)) + \lambda \sum_{i,j=1}^{n} \alpha_i \alpha_j K(X_i, X_j)$$

with the solution

$$\hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_k K(x, X_i).$$

Aggregated learning

- Try to take advantages of different classifiers.
- Boosting weak learning methods.
- The methods include model average, stacking, and boosting.

Model selection in statistical learning

- All learning methods assume f from some models.
- The choice of models is important: underfitting or overfitting.
- Often reflected in some tuning parameters in learning methods: k-NN, bandwidth, the number of basis functions, tree size, penalty parameters.
- The model selection aims to balance fitting data and model complexity.

AIC and BIC

- They apply when the loss function is the log-likelihood function and models are parametric.
- AIC: -2log-lik+2 # parameters
  BIC: -2log-lik+2log n # parameters
- Whether AIC or BIC?

Model complexity

- Not all the models have finite number of parameters.
- A more general measurement for model complexity is VC-dimension.
- Stochastic errors between the empirical risk and the limiting risk can be controlled in term of VC-dimension.
- Thus, among a series of models $\Omega_1, \Omega_2, ...$, we choose the one minimizing

$$\gamma_n(\hat{f}_\Omega) + b_n(\Omega).$$

- $\gamma_n(\hat{f}_\Omega)$ reflects the best approximation using model Ω (bias).
- $b_n(\Omega)$ is an upper bound controlling stochastic errors (variability).
- Limitation: VC-dimension is often not easy to calculate.

Cross-validation

- It is the most commonly used method.
- It is computationally feasible, although intensive.
- The idea is to use one data as training data and the other part as testing data to assess prediction error of one learning method.
- It avoids overfitting due to using only one data set
- Leave-one-out cross validation or k-fold cross-validation is used.
- Sometimes, it can be calculated quickly.

Unsupervised learning

- We don't have outcome labels but only feature data.
- We wish to see the structures within feature data.
- Useful for data exploration and dimension reduction.

Principal component analysis

- It is one popular method viewing intrinsic structure of X.
- The goal is to determine orthogonal PCs which explain most of data variations.
- It relies on singular value decomposition (SVD).

Latent component analysis

- Assume

$$X = AS + \epsilon.$$

- S are latent variables and often assumed independent from Gaussian distributions (factor analysis).
- Estimation of A is via maximum likelihood estimation.
- S can be assumed to be independent but not normally distributed (independence component analysis).

Multidimensional scaling

- This method projects original X to a much lower-dimensional space.
- It is useful for viewing X.
- The goal of the projection is to ensure pairwise distances before and after projections to be consistent as much as possible.
- Minimize

$$\left[\sum_{i \neq j}(d(X_i, X_j) - \|Z_i - Z_j\|)^2\right]^{1/2}.$$

- Can be modified to add weights to each pair or just keep distance ranks to be consistent.

Cluster analysis

- Search for clusters of subjects so that within-cluster subjects are most similar but between-cluster subjects are most different.
- Look for a map: $\mathcal{C} : \{1, ..., n\} - - > \{1, ..., K\}$ from subject ID to cluster ID.
- Within-cluster distance (loss):

$$\frac{1}{2} \sum_{i,j=1}^{n} \sum_{k=1}^{K} I(\mathcal{C}(i) = \mathcal{C}(j) = k) d(X_i, X_j).$$

- Between-cluster distance (loss):

$$\frac{1}{2} \sum_{i,j=1}^{n} \sum_{k=1}^{K} I(\mathcal{C}(i) = k, \mathcal{C}(j) \neq k) d(X_i, X_j).$$

- Either minimize within-cluster distance or maximize between-cluster distance.

K-means cluster analysis

- Applies when the distance is the Euclidean distance.
- The within–cluster distance is equivalent to

$$\sum_{i=1}^{n}\sum_{k=1}^{K} I(\mathcal{C}(i)=k)\|X_i - m_k\|^2,$$

  where m_k is the k-cluster mean.
- An iterative procedure is used to update m_k and cluster membership.

K-medoids cluster analysis

- It applies to general proximity matrix.
- Replace mean m_k by the point X_i (medoid) in the same cluster which has the least summed distance from the other points in the cluster.
- Iteratively update the medoid and cluster membership.

Hierarchical clustering

- Either agglomerative (bottom-up) or divisive (top-down).
- At each level, either merge two clusters or split clusters in an optimal sense.
- The way of defining between-cluster distance includes single linkage, complete linkage and group average.
- The output is called a dendrogram.

Bayes error in learning theory

- The classification error from the most desirable classifier:

$$\eta(X) = P(Y = 1|X),$$

$$P(I(\eta(X) > 1/2) \neq Y) = E[\min(\eta(X), 1 - \eta(X))]$$
$$= \frac{1}{2} - \frac{1}{2}E[|1 - 2\eta(X)|].$$

- Other definitions of classification errors: Komogorov variational distance, Bhattacharyya measure of affinity, Shannon entropy, Kullback-Leibler divergence.
- These errors are closely related to Bayes error.

Consistency

- Consistency of a classifier g_n (corresponding to decision function $\eta_n(x)$):

$$P(g_n(X) \neq Y) \to \text{Bayes error.}$$

- Strongly consistent:

$$P(g_n(X) \neq Y|\text{data}) \to_{a.s.} \text{ Bayes error.}$$

- Universally (strongly) consistent if the above consistency is true for any distribution of (X, Y).

A key inequality

- A key inequality:

$$P(g_n(X) \neq Y|\text{data}) \leq 2E[|\eta_n(X) - \eta(X)||\text{data}]$$

$$\leq 2E\left[(\eta_n(X) - \eta(x))^2|\text{data}\right]^{1/2}.$$

- The consistency of classifiers can be proved by showing the L_1- or L_2-consistency of η_n.

Consistency in direct learning

- It uses the key inequality.
- Since η_n often has explicit expression in direct learning, the consistency follows from the L_1- or L_2- consistency of η_n.
- For strongly consistency proof, it relies the use of concentration inequalities to conclude

$$P\left(\left|E_n[|\eta_n(X)-\eta(X)|]-E[|\eta_n(X)-\eta(X)|]\right| > \epsilon\right) \leq ae^{-nb\epsilon^2}$$

then the consistency follows from the first Borel-Cantelli lemma.

Summary of consistency results

- If the bin width $h_n \to 0$ and $nh_n^d \to \infty$, then the histogram rule is universally and strongly consistent.
- For fixed odd k, k-NN is universally consistent for the nearest neighborhood error.
- For $k \to \infty$ and $k/n \to 0$, k-NN is universally and strongly consistent.
- If the bandwidth $h \to 0$ and $nh^d \to \infty$, then the kernel rule is universally and strongly consistent.
- If the number of basis function $K_n \to \infty$ and $K_n/n \to 0$, the sieve rule is consistent and is strongly consistent if $K_n \log n/n \to 0$.

Consistency in indirect learning

- The decision rule is not explicit.
- However, we know that best classifiers minimizes some loss function or regularized loss functions.
- Thus,

$$P(L(g_n) - L(g^*) > \epsilon) \leq P(L(g_n) - L_n(g_n) - L(g^*) + L_n(g^*) > \epsilon)$$

$$\leq 2P(\sup_{g \in \mathcal{F}} |L_n(g) - L(g)| > \epsilon/2).$$

- We need control stochastic errors of such loss functions over the model space,

$$\sup_{g \in \mathcal{F}} |L_n(g) - L(g)|.$$

- This uses concentration inequalities from empirical processes and relies on the model size of $\mathcal{F}$.

Some results

- If $N(\epsilon, \mathcal{F}, L_1(P))$ is finite, then the rule based on maximum likelihood method is strongly consistent.
- If $\mathcal{F}$ has a finite VC-dimension, then the rule minimizing empirical risk

$$\sum_{i=1}^{n} I(Y_i \neq g(X_i))$$

  is strongly consistent.
- Let $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$ each having a finite VC-dimension v_k, then the rule minimizing structural risk

$$\sum_{i=1}^{n} I(Y_i \neq g(X_i)) + \sqrt{32 v_k \log(en)/n}$$

  is universally and strongly consistent if

$$\sum_{k} e^{-v_k} < \infty.$$