1. Model Selection

a) Cross Validation

- Split the data set $\mathcal{D}$ into K subsets $\mathcal{D}^{(i)}$ (called folds)
- For $i = 1, \dots, K$, train $h^{(i)}$ using all data but the i-th fold ($\mathcal{D} \backslash \mathcal{D}^{(i)}$).
- Cross-validation error by averaging all validation errors $\hat{\epsilon}_{\mathcal{D}^{(i)}}(h^{(i)})$

2. KNN

a) 误差界估计 $\lim_{n\to\infty} \epsilon_{\text{KNN}}(n) \le 2\epsilon^*(1-\epsilon^*)$

b) Normalization

- Mean $\mu_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$
- Variance $\sigma_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{ij}-\mu_j)^2}$
- Normalize the feature into a new one:

$$\hat{x}_{ij} \leftarrow \frac{x_{ij}-\mu_j}{\sigma_j}$$

c) Distance

- **Cosine Distance:** $\cos(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}{\|\boldsymbol{x}_i\| \|\boldsymbol{x}_j\|}$
- **Minkowski Distance:** $D_p(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt[p]{\sum_{k=1}^{d} |x_{ik} - x_{jk}|^p}$
- **Mahalanobis Distance:** $D_M(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \boldsymbol{M} (\boldsymbol{x}_i - \boldsymbol{x}_j)}$

d) Weighted KNN

- Distance-weighted Classification:

$$\hat{y} = \operatorname{argmax}_{c \in \mathcal{Y}} \sum_{\boldsymbol{x}' \in \text{KNN}_c(\boldsymbol{x})} \frac{1}{D(\boldsymbol{x}, \boldsymbol{x}')^2}$$

- Distance-weighted Regression:

$$\hat{y} = \sum_{\boldsymbol{x}' \in \text{KNN}(\boldsymbol{x})} \frac{1}{D(\boldsymbol{x}, \boldsymbol{x}')^2} y'$$

e) Nearest Centroid Classifier

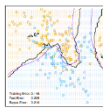
- Given dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$ with class label $y_i \in \mathcal{Y}$:
  - Compute per-class centroids $\boldsymbol{\mu}_c = \frac{1}{|N_c|}\sum_{y_i=c} \boldsymbol{x}_i$.
  - Prediction function: $\hat{y} = \operatorname{argmin}_{c \in \mathcal{Y}} \|\boldsymbol{x} - \boldsymbol{\mu}_c\|$
- $O(Nd)$ training and $O(Cd)$ testing.
  - Faster!
- Gaussian assumption for each class.
  - Cannot tackle complex distributions.
  - Very good for Few-Shot Learning (FSL).

f) Complexity: n examples, d dimensions

- For naïve KNN:
  - $O(1)$ for training.
  - $O(nd)$ to find k closest examples.
- For KNN with k-d Tree:
  - $O(dn \log n)$ for training (build k-d tree)
  - $O(2^d \log n)$ on average when query.
  - Use k-d tree only when $n \gg 2^d$.

g) Summary

- **When to use:**
  - Few attributes per instance (expensive computation)
  - Lots of training data (curse of dimensionality)
- **Advantages:**
  - Agnostically learn complex target functions
  - Do not lose information (store original data)
  - Data number can be very large (big pro!)
  - Class number can be very large (biggest pro!)
  - All other ML algorithms may fail here!
- **Disadvantages:**
  - Slow at inference time (acceleration a must)
  - Ineffective in high dimensions (curse of dimensionality)
  - Fooled easily by irrelevant attributes (feature engineering crucial)



3. Linear Regression

a) Analytic Sol

- Computing the gradient of $\hat{\epsilon}(\boldsymbol{w})$ w.r.t. $\boldsymbol{w}$ and setting it to zero yields the optimal parameter $\boldsymbol{w}^*$:

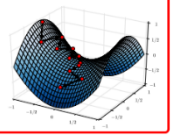
In Matrix Cookbook:
- $\|A\|_2^2 = \operatorname{tr}(A^{\mathrm{T}}A)$
- $\frac{\partial \operatorname{tr}(A^{\mathrm{T}}B)}{\partial A} = B$

$$\hat{\epsilon}(\boldsymbol{w}) = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2$$
$$\nabla_{\boldsymbol{w}} \hat{\epsilon}(\boldsymbol{w}) = 2\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) = \boldsymbol{0}$$
$$\Rightarrow \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y}$$
$$\Rightarrow \boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Normal Equation

b) Optimization: SGD

- In each iteration t ($\le T$):
  - Randomly sample a minibatch of $m \ll n$ points $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$
  - Set $J^t(\boldsymbol{w}^t) = \frac{1}{m}\sum_{i=1}^m \ell(\boldsymbol{w}^t; \boldsymbol{x}_i, y_i)$
  - Compute gradient on minibatch: $\Delta^t = \nabla_{\boldsymbol{w}} J^t(\boldsymbol{w}^t)$
  - Update parameters with learning rate η: $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta\Delta^t$



- For linear regression: $\boldsymbol{w}^{t+1} \leftarrow \boldsymbol{w}^t - 2\eta\boldsymbol{X}_m^T(\boldsymbol{X}_m\boldsymbol{w}^t - \boldsymbol{y})$
  - Computational complexity $O(dmT)$. What if large dimension d?
- Theoretically, convergence rate is $O(1/\sqrt{T})$ under convex condition.

c) Regularization

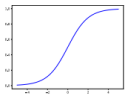
- Norm-regularization will make the hypothesis smooth at any point.

4. Linear Classification

a) Logistic Regression

- To map the output of $h(\boldsymbol{x})$ into $[0,1]$, we use sigmoid function:

$$\sigma(t) = \frac{1}{1+e^{-t}}$$



  - Sigmoid maps $\mathbb{R}$ to $[0,1]$.
  - $t \to +\infty, \sigma(t) \to 1$.
- Set $\sigma(h(\boldsymbol{x})) = p(y=1|\boldsymbol{x})$ as the probability to label $\boldsymbol{x}$ as $y = 1$.

$$\hat{\epsilon}(w) = -\sum_{i=1}^{n}\{y_i \log \sigma(h_w(\boldsymbol{x})) + (1-y_i)\log[1-\sigma(h_w(\boldsymbol{x}))]\} + \lambda\sum_{j=1}^{d} w_j^2$$

b) SoftMax Regression

- Softmax function normalizes multiple outputs in a probability vector:

$$p(y=c|\boldsymbol{x}; \boldsymbol{W}) = \frac{\exp(\boldsymbol{w}_c^T\boldsymbol{x})}{\sum_{r=1}^{C}\exp(\boldsymbol{w}_r^T\boldsymbol{x})}$$

Sum over all classes

Loss for each data point $(\boldsymbol{x}_i, y_i)$:

$$\ell(h(\boldsymbol{x}_i), y_i) = \begin{cases} -\log\frac{\exp(\boldsymbol{w}_1^T\boldsymbol{x}_i)}{\sum_{r=1}^{C}\exp(\boldsymbol{w}_r^T\boldsymbol{x}_i)}, y_i = 1 \\ -\log\frac{\exp(\boldsymbol{w}_2^T\boldsymbol{x}_i)}{\sum_{r=1}^{C}\exp(\boldsymbol{w}_r^T\boldsymbol{x}_i)}, y_i = 2 \\ \vdots \\ -\log\frac{\exp(\boldsymbol{w}_C^T\boldsymbol{x}_i)}{\sum_{r=1}^{C}\exp(\boldsymbol{w}_r^T\boldsymbol{x}_i)}, y_i = C \end{cases}$$

Equivalent to maximum log-likelihood estimation

5. SVM

a) SM SVM

Original form

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|_2^2$$
$$\text{s.t. } y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) \ge 1 - \xi_i$$
$$\xi_i \ge 0, \sum_{i=1}^{n}\xi_i \le n', 1 \le i \le n$$

Lagrangian equivalent form

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i$$
$$\text{s.t. } y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) \ge 1 - \xi_i$$
$$\xi_i \ge 0, 1 \le i \le n$$

b) Dual Problem

Primal Problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x})$$
$$\text{s.t. } g_j(\boldsymbol{x}) \le 0 \text{ for } j = 1, \dots, J$$
$$h_k(\boldsymbol{x}) = 0 \text{ for } k = 1, \dots, K$$

- With Lagrange multipliers $\boldsymbol{\lambda}, \boldsymbol{\mu}$, the Lagrangian function is defined as

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_{j=1}^{J}\lambda_j g_j(\boldsymbol{x}) + \sum_{k=1}^{K}\mu_k h_k(\boldsymbol{x})$$

KKT conditions

- KKT conditions: for $1 \le j \le J$
  - Primal feasibility: $g_j(\boldsymbol{x}) \le 0, h_k(\boldsymbol{x}) = 0$
  - Dual feasibility: $\lambda_j \ge 0$
  - Complementary slackness: $\lambda_j g_j(\boldsymbol{x}) = 0$
  - The gradient of the Lagrangian function w.r.t. $\boldsymbol{x}$ vanishes to $\boldsymbol{0}$:

$$\nabla_x f(\boldsymbol{x}) + \sum_{j=1}^{J}\lambda_j\nabla_x g_j(\boldsymbol{x}) + \sum_{k=1}^{K}\mu_k\nabla_x h_k(\boldsymbol{x}) = \boldsymbol{0}$$

- If $\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*$ satisfy KKT for a convex problem, then $\boldsymbol{x}^*$ is an **optimum**:

$$f(\boldsymbol{x}^*) = L(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$$

  - $\boldsymbol{x}^*$ is **optimal** if and only if there exist $\boldsymbol{\lambda}, \boldsymbol{\mu}$ satisfying KKT conditions.

Dual Problem

- Define the primal objective as

$$\Pi(\boldsymbol{x}) = \max_{\boldsymbol{\lambda}\in\mathbb{R}^J, \boldsymbol{\mu}\in\mathbb{R}^K} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$$

- Define the dual objective as

$$\Gamma(\boldsymbol{\lambda}, \boldsymbol{\mu}) \triangleq \min_{\boldsymbol{x}\in\mathbb{R}^d} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$$

- Weak duality: $\max_{\boldsymbol{\lambda}\in\mathbb{R}^J, \boldsymbol{\mu}\in\mathbb{R}^K} \Gamma(\boldsymbol{\lambda}, \boldsymbol{\mu}) \le \min_{\boldsymbol{x}\in\mathbb{R}^d} \Pi(\boldsymbol{x})$

Minmax theorem: $\max_x \min_y f(x,y) \le \min_y \max_x f(x,y)$

- Dual Problem:

$$\max_{\boldsymbol{\lambda}\in\mathbb{R}^J, \boldsymbol{\mu}\in\mathbb{R}^K} \Gamma(\boldsymbol{\lambda}, \boldsymbol{\mu})$$
$$\text{s.t. } \lambda_j \ge 0 \text{ for } j = 1, \dots, J$$

Soft-Margin SVM Dual Problem

- Lagrangian function (with $2n$ inequality constraints):

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu})$$
$$= \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1 - \xi_i - y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b)) - \sum_{i=1}^{n}\mu_i\xi_i$$
$$\alpha_i \ge 0, \mu_i \ge 0, i = 1, \dots, n$$

  - After solving for $\boldsymbol{\alpha}$, we can solve for $\boldsymbol{w} = \sum_{i=1}^n \alpha_i y_i \boldsymbol{x}_i$.

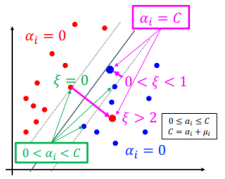
- $-\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{0} \Rightarrow \boldsymbol{w} = \sum_{i=1}^n \alpha_i y_i \boldsymbol{x}_i$
- $-\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^n \alpha_i y_i = 0$
- $-\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow C = \alpha_i + \mu_i, i = 1, \dots, n$

- Dual Problem of Soft-SVM:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j y_i y_j (\boldsymbol{x}_i \cdot \boldsymbol{x}_j)$$
$$\text{s.t. } \sum_{i=1}^n \alpha_i y_i = 0$$
$$0 \le \alpha_i \le C, 1 \le i \le n.$$

- KKT Conditions for Soft-SVM:

$$\begin{cases} \alpha_i \ge 0, \quad \mu_i \ge 0 \\ y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) - 1 + \xi_i \ge 0 \\ \alpha_i(y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) - 1 + \xi_i) = 0 \\ \xi_i \ge 0, \quad \mu_i\xi_i = 0 \end{cases}$$



- $\alpha_i(y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) - 1 + \xi_i) = 0$
  - $\alpha_i \ne 0$: Support vectors (SVs)
    - All data points that satisfy: $y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) = 1 - \xi_i \le 1$.
  - $\boldsymbol{w} = \sum_{i=1}^n \alpha_i y_i \boldsymbol{x}_i$ is a linear combination of support vectors.

c) Primal Problem: SGD

- Theorem: Soft-SVM is equivalent to **Regularized Risk Minimization**:

$$\min_{w,b} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \frac{C}{n}\sum_{i=1}^n \max\{0, 1 - y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b)\}$$

- $\ell_{0/1} = 1[t \le 0]$
- $\ell_{\text{hinge}} = \max\{1-t, 0\}$
- $\ell_{\exp} = \exp(-t)$
- $\ell_{\text{logistic}} = \log(1 + e^{-t})$
- where $t = y(\boldsymbol{w}\cdot\boldsymbol{x} + b)$

$$\min_{w,b} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \frac{C}{n}\sum_{i=1}^n \max\{0, 1 - y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b)\}$$

- SGD for Soft-SVM:
  - In each iteration t ($\le T$):
    - Choose i uniformly at random from $1, \dots, n$ (not in mini-batches)
    - The subgradient $\boldsymbol{g}_i^t = \begin{cases} \boldsymbol{w}^t & \text{if } y_i(\boldsymbol{w}^t\cdot\boldsymbol{x}_i + b) \ge 1 \\ \boldsymbol{w}^t - Cy_i\boldsymbol{x}_i & \text{otherwise} \end{cases}$
    - Update $\boldsymbol{w}^{t+1} \leftarrow \boldsymbol{w}^t - \eta\boldsymbol{g}^t$
  - Output $\bar{\boldsymbol{w}} = \frac{1}{T}\sum_{t=1}^T \boldsymbol{w}^t$

SGD for convex optimization
Convergence rate: $O(\frac{1}{\sqrt{T}})$ (📖 14.3)

d) Kernel SVM

- Commonly used kernel functions for vector data:

Linear	$k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{x}_1 \cdot \boldsymbol{x}_2$
Polynomial	$k(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\boldsymbol{x}_1 \cdot \boldsymbol{x}_2)^d$
Polynomial (with constant terms)	$k(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\boldsymbol{x}_1 \cdot \boldsymbol{x}_2 + 1)^d$
RBF (Gaussian kernel)	$k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left(-\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2}{2\sigma^2}\right)$
Hyperbolic Tangent	$k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \tanh(\beta\boldsymbol{x}_1 \cdot \boldsymbol{x}_2 + c)$

- From kernel functions k_1, k_2, we can construct new kernel functions:
  - $k'(\boldsymbol{x}_1, \boldsymbol{x}_2) = k_1 \otimes k_2(\boldsymbol{x}_1, \boldsymbol{x}_2) = k_1(\boldsymbol{x}_1, \boldsymbol{x}_2)k_2(\boldsymbol{x}_1, \boldsymbol{x}_2)$
  - For any function $g: \mathcal{X} \to \mathbb{R}$, $k'(\boldsymbol{x}_1, \boldsymbol{x}_2) = g(\boldsymbol{x}_1)k_1(\boldsymbol{x}_1, \boldsymbol{x}_2)g(\boldsymbol{x}_2)$
- Theorem (Mercer): If $k(\cdot,\cdot)$ is a symmetric function on space $\mathcal{X}\times\mathcal{X}$, then k is a kernel function $\Leftrightarrow$ For any input set $(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_m)$,

$$\boldsymbol{K} = \begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_m) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_1, \boldsymbol{x}_m) & \cdots & k(\boldsymbol{x}_m, \boldsymbol{x}_m) \end{pmatrix}$$

The kernel matrix is semi-definite ($\boldsymbol{K} \succcurlyeq 0$, i.e. $\boldsymbol{x}^T\boldsymbol{K}\boldsymbol{x} \ge 0, \forall \boldsymbol{x}$).

- For kernel functions $k_1, k_2, \dots, k_s$, and $\gamma_1, \gamma_2, \dots, \gamma_s > 0$:
  - $\sum_{i=1}^s \gamma_s k_s$ is also (multi-)kernel function, because $\sum_{i=1}^s \gamma_s \boldsymbol{K}_s \succcurlyeq 0$.
- Dual Problem of Soft-SVM:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j)$$
$$\text{s.t. } \sum_{i=1}^n \alpha_i y_i = 0$$
$$0 \le \alpha_i \le C, 1 \le i \le n.$$

SS

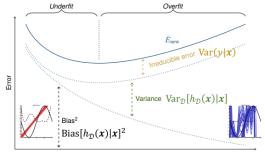
  - After solving for $\boldsymbol{\alpha}$, we can solve for $\boldsymbol{w} = \sum_{i=1}^n \alpha_i y_i \boldsymbol{\Phi}(\boldsymbol{x}_i)$.
  - Testing: $f(\boldsymbol{x}) = \boldsymbol{w}\cdot\boldsymbol{\Phi}(\boldsymbol{x}) + b = \sum_{i=1}^n \alpha_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}) + b$.
  - During training and testing, only need to compute $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

SS

6. Learning Theory

a) Bias-Variance Decomposition

- **Theorem:** $\mathcal{E}_{L2}(\boldsymbol{x}) = \text{Var}(y|\boldsymbol{x}) + \text{Bias}[h_{\mathcal{D}}(\boldsymbol{x})|\boldsymbol{x}]^2 + \text{Var}_{\mathcal{D}}[h_{\mathcal{D}}(\boldsymbol{x})|\boldsymbol{x}]$.



b) Approximation Error and Estimation Error

- General equality: given a target function f, for any $h \in \mathcal{H}$,

$$\mathcal{E}(h) - \mathcal{E}^*(f) = \underbrace{[\mathcal{E}(h) - \mathcal{E}(h^*)]}_{\text{estimation}} + \underbrace{[\mathcal{E}(h^*) - \mathcal{E}^*(f)]}_{\text{approximation}}$$

c) PAC Learning

- A hypothesis space $\mathcal{H}$ is PAC-learnable if there exists an algorithm $\mathcal{A}$ and a polynomial function $\mathbf{poly}()$, such that for any $\varepsilon > 0, \delta > 0$, for all distributions D on $\mathcal{X}$ and for any target hypothesis $h \in \mathcal{H}$, the following holds for sample complexity $n \geq \mathbf{poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, |\mathcal{H}|)$:

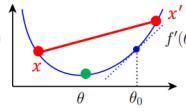
$$P_{\mathcal{D}_n \sim D^n}\left[\mathcal{E}(h_{\mathcal{D}_n}) - \min_{h \in \mathcal{H}} \mathcal{E}(h) \geq \varepsilon\right] \leq \delta$$

$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, \mathcal{E}(h)$ — Approximately correct — Probably correct

d) Probability Tools

Basic

- Union bound: $P(A \vee B) \leq P(A) + P(B)$.
- Inversion: if $P(X \geq \epsilon) \leq f(\epsilon)$, then for any $\delta > 0$, with probability at least $1 - \delta$, $X \leq f^{-1}(\delta)$.
  - Useful to interchange ϵ and δ: $P\left(\left|\hat{\mathcal{E}}_{\mathcal{D}_n}(h) - \mathcal{E}(h)\right| \geq \varepsilon\right) \leq \delta$.
- Jensen' inequality: if f is convex, then $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$.
  - Convex function: f is convex if $\forall \lambda \in [0,1], x, x' \in \mathbf{dom}(f)$:
    - $f(\lambda x + (1-\lambda)x')$ $\leq \lambda f(x) + (1-\lambda) f(x')$
    - $f(x') \geq f(x) + g^T(x' - x)$
      - g: Subgradient.



Concentration Inequalities

Markov's Inequality

Theorem: If Z is a nonnegative random variable, then for any $\varepsilon > 0$:

$$P(Z \geq \varepsilon) \leq \frac{\mathbb{E}[Z]}{\varepsilon}$$

Proof:

Chebyshev's Inequality

$$P(V \geq \varepsilon) \leq P(\phi(V) \geq \phi(\varepsilon)) \leq \frac{\mathbb{E}[\phi(V)]}{\phi(\varepsilon)}$$

Chernoff Bound

- Theorem: If we take $\phi(t) = e^{\lambda t}$, we have for any $\lambda > 0$:

$$P(V \geq \varepsilon) \leq \frac{\mathbb{E}\left[e^{\lambda V}\right]}{e^{\lambda \varepsilon}}$$

- If we take $V = Z - \mathbb{E}[Z]$, we have:

$$P(Z - \mathbb{E}[Z] \geq \varepsilon) \leq \frac{\mathbb{E}\left[e^{\lambda(Z - \mathbb{E}[Z])}\right]}{e^{\lambda \varepsilon}}$$

- If we set Z as the sum of i.i.d. random variables: $Z = \sum_{i=1}^n X_i$

$$P\left(\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq \varepsilon\right) \leq \frac{\mathbb{E}\left[e^{\lambda \sum_{i=1}^n (X_i - \mathbb{E}X_i)}\right]}{e^{\lambda \varepsilon}} = \frac{\prod_{i=1}^n \mathbb{E}\left[e^{\lambda (X_i - \mathbb{E}X_i)}\right]}{e^{\lambda \varepsilon}}$$

Hoeffding's Lemma

- Lemma: Let V be a bounded random variable, $\mathbb{E}[V] = 0$ and $a \leq V \leq b$ with $b > a$. Then for any $\lambda > 0$:

$$\mathbb{E}\left[e^{\lambda V}\right] \leq e^{\frac{\lambda^2 (b-a)^2}{8}}$$

- Plug in the optimal $\lambda = 4\varepsilon / \sum_{i=1}^n (b_i - a_i)^2 \Rightarrow$ Hoeffding's inequality:

$$P\left(\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq \varepsilon\right) \leq \frac{\prod_{i=1}^n \mathbb{E}\left[e^{\lambda (X_i - \mathbb{E}X_i)}\right]}{e^{\lambda \varepsilon}} \leq e^{\sum_{i=1}^n \frac{\lambda^2 (b_i - a_i)^2}{8} - \lambda \varepsilon} \overset{\min}{\Longrightarrow} \exp\left[-\frac{2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right]$$

Chernoff Bound — Hoeffding's Lemma — Optimal λ

- If we replace X_i with $-X_i \in [-b_i, -a_i]$, we have

$$P\left(\sum_{i=1}^n (X_i - \mathbb{E}X_i) \leq -\varepsilon\right) \leq \exp\left[-\frac{2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right]$$

- Theorem (Hoeffding's Inequality): for $a_i \leq X_i \leq b_i$, finite-sample bound

$$P\left(\left|\sum_{i=1}^n (X_i - \mathbb{E}X_i)\right| \geq \varepsilon\right) \leq 2\exp\left[-\frac{2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right]$$

McDiarmid's Inequality

- Theorem: Let $X_1, X_2, \ldots, X_n \in \mathcal{X}$ be independent random variables and $f: \mathcal{X}^n \to \mathbb{R}$ be a function that is stable to input change:

$$\sup_{x_1, x_2, \ldots, x_n, x_i'} |f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i$$

for all $i \in [1, n]$. Then, for all $\varepsilon > 0$: finite-sample bound!

$$P(|f(X_1, \ldots, X_n) - \mathbb{E}f(X_1, \ldots, X_n)| \geq \varepsilon) \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

e) Generalization Bound: Finite Hypothesis Space

- Theorem: Let $\mathcal{H}$ be a finite hypothesis space, $|\mathcal{H}| < \infty$, then for any $\delta > 0$, with probability at least $1 - \delta$,

$$\forall h \in \mathcal{H}, \quad \mathcal{E}(h) \leq \hat{\mathcal{E}}_{\mathcal{D}_n}(h) + \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2n}}$$

f) Rademacher Complexity

- Let $\mathcal{G}$ be a family of general functions mapping from $\mathcal{Z}$ to $[0,1]$.
- Let σ_i be i.i.d. uniform random variables (i.e. Rademacher variable):

$$P(\sigma_i = 1) = 1/2, P(\sigma_i = -1) = 1/2$$

- Empirical Rademacher Complexity of $\mathcal{G}$ on a size-n sample set $\mathcal{S}_n = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\}$: Average loss on all dichotomies

$$\hat{\mathcal{R}}_{\mathcal{S}_n}(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(\boldsymbol{z}_i)\right]$$

Peter Bartl

- (Expected) Rademacher Complexity:

$$\mathcal{R}_n(\mathcal{G}) = \mathbb{E}_{\mathcal{S}_n \sim D^n} \hat{\mathcal{R}}_{\mathcal{S}_n}(\mathcal{G}) = \mathbb{E}_{\mathcal{S}_n \sim D^n} \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(\boldsymbol{z}_i)\right]$$

- Theorem: Rademacher complexity will decrease when n increases:

$$\mathcal{R}_{n+1}(\mathcal{G}) \leq \mathcal{R}_n(\mathcal{G})$$

- Theorem: Let $\mathcal{H}$ be a family of binary classifiers taking values in $\{-1, +1\}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the generalization bound holds for all $h \in \mathcal{H}$:

$$\mathcal{E}_D(h) \leq \hat{\mathcal{E}}_{\mathcal{D}_n}(h) + \mathcal{R}_n(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

$$\mathcal{E}_D(h) \leq \hat{\mathcal{E}}_{\mathcal{D}_n}(h) + \hat{\mathcal{R}}_n(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

g) VC dim

- VC-dimension of a hypothesis space $\mathcal{H}$ is defined by

$$\mathbf{VCdim}(\mathcal{H}) = \max\{n : \Pi_{\mathcal{H}}(n) = 2^n\}$$

  - VC-dimension is essentially the size of the largest set that can be fully shattered by $\mathcal{H}$.
    - No need to shatter every sample set of size n.
  - Lower bound: Build a sample of size n that can be shattered.
  - Upper bound: Prove all samples of size $n+1$ cannot be shattered.

- Let $\mathcal{H}$ be a hypothesis set with $\mathbf{VCdim}(\mathcal{H}) = d$
  - If $d = \infty, \Pi_{\mathcal{H}}(n) \leq 2^n$. — $d < \infty$ is PAC learnable!
  - If $d < \infty, \Pi_{\mathcal{H}}(n) \leq \left(\frac{en}{d}\right)^d = O(n^d)$.

Vladimir Vapnik
Statistical learning theory

- Theorem: Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$ with VC-dimension d, then for any $\delta > 0$, with probability at least $1 - \delta$, for all $h \in \mathcal{H}$:

$$\mathcal{E}_D(h) \leq \hat{\mathcal{E}}_{\mathcal{D}_n}(h) + \sqrt{\frac{2d\log\frac{en}{d}}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

- The general form: $\mathcal{E}_D(h) \leq \hat{\mathcal{E}}_{\mathcal{D}_n}(h) + O\left(\sqrt{\frac{\log(n/d)}{(n/d)}}\right)$.

7. Decision Tree

- Misclassification error:

$$\mathrm{Err}(\mathcal{D}) = 1 - \max_{1 \leq k \leq K}\left(\frac{|\mathcal{C}_k|}{|\mathcal{D}|}\right)$$

- Entropy (used in ID3 and C4.5):

$$H(\mathcal{D}) = -\sum_{k=1}^K \frac{|\mathcal{C}_k|}{|\mathcal{D}|} \log\frac{|\mathcal{C}_k|}{|\mathcal{D}|}$$

- Gini index (used in CART):

$$\mathrm{Gini}(\mathcal{D}) = 1 - \sum_{k=1}^K \left(\frac{|\mathcal{C}_k|}{|\mathcal{D}|}\right)^2$$

a) ID3

Class label

ID3(*Examples*, *Target_attribute*, *Attributes*)

- create a *Root* node for the tree; assign all *Examples* to *Root*; — Stop Criteria
- if all *Examples* are positive, return the single-node tree *Root*, with label=+;
- if all *Examples* are negative, return the single-node tree *Root*, with label=−;
- if *Attributes* is empty, return the single-node tree *Root*, with label = the most common value of *Target_attribute* in *Examples*;
- otherwise // Main loop:
  - $A \leftarrow$ the attribute from *Attributes* that best* classifies *Examples*; — $O(dn)$ in each layer
  - the decision attribute for *Root* $\leftarrow A$;
  - for each possible value v_i of A
    - add a new tree branch below *Root*, corresponding to the test $A = v_i$;
    - let $Examples_{v_i}$ be the subset of *Examples* that have the value v_i for A;
    - if $Examples_{v_i}$ is empty
      - below this new branch add a leaf node with label = the most common value of *Target_attribute* in *Examples*;
    - else
      - below this new branch add the subtree $\mathrm{ID3}(Examples_{v_i}, Target\_attribute, Attributes \backslash \{A\})$; — depth $\min(d, \log n)$
- return *Root*;

* The *best* attribute is the one with the highest *information gain*.

b) C4.5

IG Rate

- By penalizing multivalued rate, IG is improved to Gain Ratio (GR):

$$\mathrm{GR} = \frac{\text{Information Gain}}{\text{multivalued rate}}$$

- C4.5 Algorithm measures multivalued rate by Intrinsic Value (IV):

$$\mathrm{IV}(f) = -\sum_{i=1}^{|V|} \frac{|\mathcal{D}_i|}{|\mathcal{D}|} \log\frac{|\mathcal{D}_i|}{|\mathcal{D}|}$$

  - When $\mathcal{D}$ is split by the feature f to $\mathcal{D}_i$ if feature value is i.
  - Which is the entropy of the value probability of the feature f.

Attribute with costs

- C4.5 introduces a new criterion that could penalize the cost:

$$\frac{(\text{Gain Ratio})^2}{\text{Cost}}$$

Missing Value

- Step 1: When computing IG, for a feature with missing values:
  - Compute the ratio of samples with missing value: $|\bar{\mathcal{D}}|/|\mathcal{D}| = \rho$.
  - Compute the IG on $\mathcal{D} \backslash \bar{\mathcal{D}}$ (all samples with values):
  - The IG for this feature is computed by: $(1 - \rho) \cdot$ IG.
    - Then compute Gain Ratio based on this new Information Gain.

Continuous Variables

- If $x_{1j}, \ldots, x_{nj}$ are the sorted values of the jth feature for n instances.
  - We only need to check split points between adjacent values
  - Traditionally take split points halfway between adjacent values:

$$s_j \in \left\{\frac{1}{2}(x_{rj} + x_{(r+1)j}) \middle| r = 1, \ldots, n-1\right\}$$

- Check performance of $n-1$ splits and find the one with highest IG.

Pruning

- Pre-Pruning
  - If the accuracy grows no more than eta or even decreases: Stop the split
  - May cause underfitting
- Post-Pruning
  - Avoid underfitting

- The cost complexity criterion with parameter α for tree T is

$$C_\alpha(T) = \hat{\mathcal{E}}(T) + \alpha|T| = \sum_{t=1}^{|T|} N_t H_t(T) + \alpha|T|$$

$$= -\sum_{t=1}^{|T|} \sum_{k=1}^K N_{tk} \log\frac{N_{tk}}{N_t} + \alpha|T|$$

  - $|T|$: #leaf nodes, N_{tk} is the #examples of class k in leaf node t
  - $\hat{\mathcal{E}}(T)$ is the empirical error of the tree on training data.
- Cost Complexity Pruning: (Given tree T_0 and parameter α)
  - Compute the empirical entropy $H_t(T)$ of each node
  - Recursively shrink from leaf nodes to internal nodes
    - If $C_\alpha(T_B) \leq C_\alpha(T_A)$: prune leaf A and use parent B as new leaf

c) CART

Classification Tree

- If $\mathcal{D}$ is split into $\mathcal{D}_1$ and $\mathcal{D}_2$ by whether attribute $A(x) = a$:

$$\mathcal{D}_1 = \{(x, y) \in \mathcal{D} : A(x) = a\}, \mathcal{D}_2 = \mathcal{D} - \mathcal{D}_1$$

- Then under the condition of attribute A, the Gini index of $\mathcal{D}$ is

$$\mathrm{Gini}(\mathcal{D}, A) = \frac{|\mathcal{D}_1|}{|\mathcal{D}|}\mathrm{Gini}(\mathcal{D}_1) + \frac{|\mathcal{D}_2|}{|\mathcal{D}|}\mathrm{Gini}(\mathcal{D}_2)$$

- Gini index $\mathrm{Gini}(\mathcal{D}, A)$ is used to evaluate node split in CART, which is a binary tree where each node is split into $A(x) = a$ and $A(x) \neq a$

Regression Tree

- For each splitting variable j and splitting point s
  - Assume that the region R is split by j and s to R_1 and R_2.

$$\hat{v}_1(j,s) = \mathrm{avg}(y_i | \boldsymbol{x}_i \in R_1(j,s))$$

$$\hat{v}_2(j,s) = \mathrm{avg}(y_i | \boldsymbol{x}_i \in R_2(j,s))$$

- Find j, s that minimize the loss

$$\ell(j,s) = \underbrace{\sum_{i: \boldsymbol{x}_i \in R_1(j,s)} (y_i - \hat{v}_1(j,s))^2}_{\text{L2 loss for samples in } R_1(j,s)} + \underbrace{\sum_{i: \boldsymbol{x}_i \in R_2(j,s)} (y_i - \hat{v}_2(j,s))^2}_{\text{L2 loss for samples in } R_2(j,s)}$$

- As we did in classification, search on all (j, s)'s and find a best one.

8. Random Forest

a) Bagging

- Bagging is a general technique to reduce the variance of an estimator.
- Draw B bootstrap samples $\mathcal{D}_n^1, \ldots, \mathcal{D}_n^B$ from original data $\mathcal{D}_n$.
- Let $h_{\mathcal{D}_n^1}, \ldots, h_{\mathcal{D}_n^B}$ be the prediction functions for each sample.
- The bagged prediction function is a combination of these functions:

$$h_{\mathrm{bag}}(x) = \mathrm{Combine}\left(h_{\mathcal{D}_n^1}(\boldsymbol{x}), \ldots, h_{\mathcal{D}_n^B}(\boldsymbol{x})\right)$$

- How can we combine:
  - Binary/multiclass predictions, Class probability predictions — Vote — $\mathrm{Var}(h_{\mathrm{bag}}) \leq \mathrm{Var}(h_{\mathcal{D}_n})$
  - Prediction functions for regression — Average

Leo Breiman

- For each training point $\boldsymbol{x}_i$, let

$$S_i = \{b | \mathcal{D}_n^b \text{ does not contain } \boldsymbol{x}_i\}$$

- The OOB prediction on $\boldsymbol{x}_i$ is

$$h_{\mathrm{OOB}}(\boldsymbol{x}_i) = \frac{1}{|S_i|} \sum_{b \in S_i} h_{\mathcal{D}_n^b}(\boldsymbol{x}_i)$$

b) Breiman Algorithm

Input: Data set $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$;
Feature subset size K.

Process:

1. $N \leftarrow$ create a tree node based on D;
2. **if** *all instances in the same class* **then return** N
3. $\mathcal{F} \leftarrow$ the set of features that can be split further;
4. **if** $\mathcal{F}$ *is empty* **then return** N
5. $\tilde{\mathcal{F}} \leftarrow$ select K features from $\mathcal{F}$ randomly;
6. $N.f \leftarrow$ the feature which has the best split point in $\tilde{\mathcal{F}}$;
7. $N.p \leftarrow$ the best split point on $N.f$;
8. $D_l \leftarrow$ subset of D with values on $N.f$ smaller than $N.p$;
9. $D_r \leftarrow$ subset of D with values on $N.f$ no smaller than $N.p$;
10. $N_l \leftarrow$ call the process with parameters (D_l, K);
11. $N_r \leftarrow$ call the process with parameters (D_r, K); $K \approx \sqrt{d}$
12. **return** N

Randomized features

Output: A random decision tree

each tree