Linear Regression.

Legend: 类别 概率 最优 公式 损失 优点

- Prediction : $\hat{y} = x^T\hat{\beta}$.
- Cost : $J(\beta) = \frac{1}{2}\sum(x_i^T\beta - y_i)^2/n$ Ordinary Lease Square

  $J(\beta) = (X^T\beta - y)^T(X^T\beta - y)/2n$

  $\frac{\partial J}{\partial \beta} = (X^TX\beta - X^Ty)/n = 0$

  $\hat{\beta} = (X^TX)^{-1}X^Ty$
- Batch GD: For $\{\beta^{t+1} = \beta^t - \eta\frac{\partial J}{\partial \beta}\big|_{\beta=\beta^t}\}$

  $\frac{\partial J}{\partial \beta} = \sum\frac{\partial J_i}{\partial \beta}/n = \sum x_i(x_i^T\beta - y_i)/n$

Normalization makes every feature equally important, and prevents fast descent (bad for GD)

- Stochastic GD:

  For{ For $i=1:n$ $\{\beta^{t+1} = \beta^t + \eta(y_i - x_i^T\beta^t)x_i\}$ }
- $x \sim N(\mu, \sigma^2) \Rightarrow f(x=x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- Model $y_i = x_i^T\beta + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$

  $y_i | x_i, \beta \sim N(x_i^T\beta, \sigma^2)$, $E(y_i|x_i) = x_i^T\beta$

  Likelihood: $L(\beta) = \prod_i p(y_i|x_i, \beta)$

  $= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i - x_i^T\beta)^2}{2\sigma^2}\right)$
- L2 Regularization : Add λI ⟵ If X^TX not invertible

  $\sum(y_i - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$
- MSE: Mean Square Error

  $\frac{J(\hat{\beta})}{n} = \frac{\sum(x_i^T\hat{\beta} - y_i)^2}{n} \to E[(\hat{f}(x) - f(x) - \varepsilon)^2] = \text{bias}^2 + \text{variance} + \varepsilon$.

Logistic Regression.

- Discrete value $P(Y = C_j|x, \beta) \in [0,1]$, $\sum P(Y = C_j|x, \beta) = 1$. 类别, 比如 Y=0 和 Y=1.
- Logistic/Sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$

  $P(Y=1|x, \beta) = \sigma(x^T\beta) = \frac{\exp(x^T\beta)}{1+\exp(x^T\beta)}$

  $P(Y=0|x, \beta) = 1-\sigma(x^T\beta) = \frac{1}{1+\exp(x^T\beta)}$

  $y|x, \beta \sim \text{Bernoulli}(\sigma(x^T\beta))$
- $L = \prod_i p_i^{y_i}(1-p_i)^{y_i} = \prod\left(\frac{\exp(x_i^T\beta)}{1+\exp(x_i^T\beta)}\right)^{y_i}\left(\frac{1}{1+\exp(x_i^T\beta)}\right)^{1-y_i}$
- $\log L = \sum y_i x_i^T\beta - \log(1+\exp(x_i^T\beta))$ $\xrightarrow{-\beta_j\text{偏导}}$ $\frac{x^T\cdot\exp(x^T\beta)}{1+\exp(x^T\beta)} = P_i(\beta)x_{ij}$
- Gradient Ascent: $\beta^{new} = \beta^{old} + \eta\frac{\partial \log L}{\partial \beta}$
- Newton-Raphson: $\beta^{new} = \beta^{old} - \left(\frac{\partial^2 \log L}{\partial\beta\partial\beta^T}\right)^{-1}\frac{\partial \log L}{\partial \beta}$
- 一阶导: $\frac{\partial \log L(\beta)}{\partial \beta_j} = \sum x_{ij}(y_i - P_i(\beta))$, for $j = 0, 1, \cdots, p$
- 二阶导: $\frac{\partial \log L(\beta)}{\partial \beta_j \partial \beta_n} = -\sum_{i=1}^{N} x_{ij}x_{in}P_i(\beta)(1-P_i(\beta))$

P19
- Cross Entropy loss $H(q,p) = -\sum q_k\log(p_k)$.

  $\Rightarrow H(q,p) = -yx^T\beta + \log(1+\exp(x^T\beta))$
- Decision boundary is linear $\sigma(x^T\beta) \geq a.5 \Rightarrow x^T\beta \geq 0$

Decision Tree

- $H(y) = -\sum q_k\log(p_k)$. $p_i = P(Y=y_i)$
- $\text{Info}(D) = -\sum P_i\log_2(P_i)$
- Conditional Gain : $\text{Info}_A(D) = \sum\frac{|D_j|}{|D|}\times\text{Info}(D_j)$
- $\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$
- Algorithm. Stop Condition ① Node's samples in 1 class

  ② No Remained Attributes → Majority Voting. ③ No sample left. → MV parent.
- Split Info A (D) $= -\sum\frac{|D_j|}{|D|}\cdot\log_2\left(\frac{|D_j|}{|D|}\right)$ → 倾向分多的属性 bias to attribute with large number of values

  Gain Ratio (A) = Gain (A) / Split Info (A) 倾向unbalanced splits
- Gini Index. $\text{Gini}(D) = 1 - \sum P_j^2$, $\text{Gini}_A(D) = \frac{|D_1|}{|D|}\text{gini}(D_1) + \frac{|D_2|}{|D|}\text{gini}(D_2)$
- 防Overfitting : Prepruning, Postpruning

P4-14
- Random Forest. $p' = \sqrt{p}$, Accuracy↑ large Data.v.
- ID3 not always optimal, eg. XOR in 18mid (min err & path)

SVM

Right for next same point: $yw^Tx = yw^Tx + \eta y^2\|x\|^2 > 0$

- Time Consuming / Compute distance between any 2 points
- Point to Plane: $\left|(x_0-x, y_0-y, z_0-z)\cdot\frac{(a,b,c)}{\|(a,b,c)\|}\right| = \frac{|ax_0+by_0+cz_0-d|}{(a^2+b^2+c^2)^{\frac{1}{2}}}$
- Simple Perceptron : $x = (1, x_1, \cdots x_d)^T$. If $y \neq \hat{y}$: $w = w + \alpha(y_i x_i)$
- Support vectors: Pts on margin hyperplanes H_1, H_2
- $\rho = \frac{2}{\|w\|}$ → minimize $\Phi(w) = \frac{1}{2}w^Tw$, st. $y_i(w^Tx_i+b) \geq 1$ primary
- Lagrangian: $L(w, \alpha, \beta) = f(w) + \sum\alpha_i h_i(w) + \sum\beta_j g_j(w)$ $g_j(x)\leq 0$
- KKT ① L 对任何 w 的梯度 = 0, ② $h(x) = 0$ ③ $\sum\alpha_j g_j(x) = 0$, $\alpha_j \geq 0$
- $L(w, b, \alpha) = \frac{1}{2}w^Tw - \sum\alpha_i(y_i(w^Tx_i+b) - 1)$

  $\nabla_w L = 0 \Rightarrow w = \sum_{i=1}^{N}\alpha_i y_i x_i \Rightarrow \alpha > 0$ 时为 SV.

  $\frac{\partial L}{\partial b} = 0 \Rightarrow -\sum\alpha_i y_i = 0$
- Dual : Find 参数 α st. Max $Q(\alpha) = -\frac{1}{2}\sum_i\sum_j\alpha_i\alpha_j y_i y_j x_i^Tx_j$, st. $\sum\alpha_i y_i = 0$, $\alpha_i \geq 0$ for $\forall\alpha_i$ 内积, needed.
- Solution : $w = \sum\alpha_i y_i x_i$, $b = y_k - w^Tx_k$ $\Rightarrow f(x) = \sum\alpha_i y_i x_i^Tx + b$
- Soft Margin : $\min_{\lambda}\Phi(w,w) = \frac{1}{2}w^Tw + C\sum\varepsilon_i$ st $y_i(w^Tx_i+b) \geq 1-\varepsilon_i$, $\varepsilon_i \geq 0$ for all i ↳ Regularization Optimize

  ⇒ Dual Problem 不变, 加 "$0 < \alpha_i < C$ 时 $\varepsilon_i = 0$; $\alpha_i = C$ 时 $\varepsilon_i > 0$".
- Kernels : $K(x_i, x_j) = (x_i\cdot x_j + 1)^h$ $K(x_i, x_j) = e^{-\|x_i - x_j\|^2/2\sigma^2}$

  $K(x_i, x_j) = \tanh(\kappa x_i\cdot x_j - \delta)$ Eg. $K = (1+xz)^2$ $\phi = [1, \sqrt{2}x, x^2]$

  Decision boundary : $f(x) = \sum\alpha_i y_i K(x_i, x) + b$

Pro: Computes K without computes $\phi(x)$

For slack: when $C \to 0$, all points are SV. When $C \to \infty$, hard SVM.

NN

- Activation function controls output's nature, add nonlinearity, differentiable
- Loss function $L(w) = \frac{1}{n}\sum L(y, \hat{y})$, $\hat{y}^{(i)} = f(x^{(i)}, w)$ $y \in \{0,1\}$, $\hat{y} \in [0,1]$

  Square Error $L(y,\hat{y}) = (y - \hat{y})^2$, (Binary) CEL : $L(y,\hat{y}) = -y\log\hat{y} - (1-y)\log(1-\hat{y})$

  Hinge $L(y,\hat{y}) = \max(0, 1-\hat{y}y)$, $y \in \{-1,1\}$, $\hat{y} \in [-\infty, \infty]$ Exponential: $L = e^{-y w^Tx}$

  Single Neuron Perceptron : $L(y,\hat{y}) = \max(0, -y\hat{y})$, Linear SVM: $L = \max(0, 1-\hat{y}y)$ & L2

  Logistic : Sigmoid + CEL Logistic Loss: $L = \log(1+e^{-yw^Tx})$
- MLP: Multi Layer Perceptron → any continuous function
- Backpropagation : $E = \frac{1}{2}\|y - x^{(2)}\|^2 \Rightarrow \frac{\partial E}{\partial w_{j:(1)}} = \frac{\partial E}{\partial x_{i:(2)}}\frac{\partial x_{i:(2)}}{\partial z_{i:(2)}}\frac{\partial z_{i:(2)}}{\partial w_{j:(1)}} = -(y_i - x_i^{(2)})f'(z_i^{(2)})x_j^{(1)}$

KNN not linear

- Small K : overfitting (high var, low bias). Big K: Too many irrelevant Pts
- Minkowski Distance : $d(i,j) = (|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h)^{\frac{1}{h}}$

  $h=1 \to$ manhattan ; $h=2 \to$ Euclidean , $h=\infty \Rightarrow$ supremum 最远方向的距离.
- mean absolute deviation : $s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \cdots)$
- Nominal Attribute $d(i,j) = \frac{p-m}{p}$, m is matched.
- Distance for asymmetric binary $d(i,j) = \frac{r+s}{q+r+s}$

	1	0
1	q	r
0	s	t

- Ordinal Variable : Map to [0,1] $z_{if} = \frac{r_{if}-1}{M_f-1}$, $r_{if} \in \{1, \cdots, M_f\}$
- weighted formula : $d(i,j) = \frac{\sum\delta d}{\sum\delta}$
- cosine similarity : $\cos(d_1, d_2) = (d_1\cdot d_2)/\|d_1\|\cdot\|d_2\|$

Evaluations

- Holdout : Training set + Test set
- Accuracy : (TP+TN)/All. Error Rate : (FP+FN)/All

  Sensitivity : TP/P Specificity : TN/N

  Precision : TP/P' Recall : TP/TP+FN = Sensitivity = TPR = 1 - FPR

  F-score : $2/(\text{precision}^{-1} + \text{recall}^{-1})$
- ROC: Receiver Operating Characteristics. 按 likely of positive 排序, 从"左"predict "N"到"右"P".
- Multi Class : One-vs-All & All-vs-All → prefered, $O(n^2)$
- Oversampling & Undersampling

Hierarchical Clustering

- AGNES (Agglomerative Nesting)
  - Init: own cluster. compute distance matrix
  - Repeat: Merge 2 closest clusters
    Update affected entries in matrix
  - Until: 1 single cluster. 不用specify level
  - Clustering is cutting the dendrogram at certain level.

- DIANA (Divisive Analysis) Inverse of AGNES
  - Distance between clusters: Single, Complete (max), Average, Centroid, Medoid Link.

- Density-Based: Arbitrary shape, noise✓ one scan, density param.
  - DBSCAN: Eps (radius), MinPts
    density-reachable, density-connected,
  - Algo: 任意pt周围pts加进来. if Pts > MinPts.

K-Means

- Objective $J = \sum_{j=1}^{k} \sum_i w_{ij} \|x_i - c_j\|^2$
- ① Fix centers, find w_{ij}: $w_{ij} = 1$ if $\|x_i - c_j\|^2$ is the smallest
  ② Fix w_{ij}, find centers: $\frac{\partial J}{\partial c_j} = -2\sum w_{ij}(x_i - c_j) = 0$
  $\Rightarrow c_j = \frac{\sum w_{ij} x_i}{\sum w_{ij}}$
- Problems: ① different size/density ② Non-spherical ③ fix K ④ Outlier
- hard clustering, $w_{ij} = \{0, 1\}$, $\sum w_{ij} = 1$.
- k-modes for categorical data. k-Medoid for smallest total dist.
- Best K: Knee points. (Bigger K always smaller cost)

GMM

- Soft clustering $w_{ij} \in [0, 1]$. $\sum w_{ij} = 1$
- $P(x_i, z_i = c_j) = w_j f_j(x_i)$
- Likelihood: $P(x_i) = \sum_j w_j f_j(x_i)$
- MLE. objects are independent. $D = \{x_1, \dots x_n\}$.
  $p(D) = \prod_i P(x_i) = \prod_i \sum_j w_j f_j(x_i)$
  $\Rightarrow \log p(D) = \sum_i \log p(x_i) = \sum_i \log \sum_j w_j f_j(x_i)$
- EM Algorithm
  E. $w_{ij}^{t+1} = P(z_i = j \mid \theta_j^t, x_i) \propto \underbrace{p(x_i \mid z_i = j, \theta_j^t)}_{f_j(x_i)} \underbrace{P(z_i = j)}_{w_j}$
  M. $\theta^{t+1} = \arg\max_\theta \sum_i \sum_j w_{ij}^{t+1} \log p(x_i, z_i = j \mid \theta)$
- Generative Model.
  $Z \sim \text{Multinoulli}(w_1, \dots w_k)$, $X \mid Z \sim N(\mu_z, \sigma_z^2)$
  Overall Likelihood $L(D \mid \theta) = \prod_i \sum_j w_j p(x_i \mid \mu_j, \sigma_j^2)$
  E $w_{ij}^{t+1} \propto w_j^t p(x_i \mid \mu_j^t, \sigma_j^{2\,t}) \rightarrow f_j^t(x_i)$
  M Find μ_j, σ_j, w_j: $\sum_{ij} w_{ij}^{t+1} \log w_j p(x_i \mid \mu_j, \sigma_j^2)$
  When having w_{ij}, it is equivalent to Gaussian Distribution
  $\mu_j^{t+1} = \frac{\sum_i w_{ij}^{t+1} x_i}{\sum_i w_{ij}^{t+1}}$ $(\sigma_j^2)^{t+1} = \frac{\sum_i w_{ij}^{t+1}(x_i - \mu_j^{t+1})^2}{\sum_i w_{ij}^{t+1}}$, $w_j^{t+1} = \sum_i w_{ij}^{t+1} / n$
- 2D: P25
- When each component same $\sigma^2 I$ and $w_j \rightarrow$ soft k-means
- When $\sigma^2 \to 0$, hard assignment, k-means ✓.
- Mapping soft clustering → hard cluster: 选 w_{ij} 最大的为1.

Clustering Evaluation

- Extrinsic: supervised. eg. Purity, Precision & recall
- Intrinsic: unsupervised. eg. Silhouette coefficient.
- C is output. Ω is ground truth. $\{C_1, \dots C_k\}$, $\{w_1, \dots w_j\}$
- Purity $(C, \Omega) = \frac{1}{N} \sum_k \max_j |C_k \cap w_j|$ i.e. the proportion of matched pts.
- NMI $(C, \Omega) = I(C, \Omega) / \sqrt{H(C) H(\Omega)}$
- $I(\Omega, C) = \sum_k \sum_j P(C_k \cap w_j) \log \frac{P(C_k \cap w_j)}{P(C_k) \cdot P(w_j)} = \sum_k \sum_j \frac{|C_k \cap w_j|}{N} \log \frac{N |C_k \cap w_j|}{|C_k| \cdot |w_j|}$
- $H(\Omega) = -\sum_j P(w_j) \cdot \log P(w_j)$
- Random Index (RI) $= \frac{TP + TN}{All}$, F-score $= \frac{2 \text{ Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
- Clustering. 每两个pts 求TP/TF/NP/NF.
- If flip truth label evaluation not the same.
- BIC: $2 \underbrace{\ln(x, \theta)}_{\text{log likelihood}} - \underbrace{mn}_{\text{\# of params to be estimated}} \log(n) \equiv BIC$

Set Data

- Frequent Pattern
- Itemset (>0), k-itemset, (absolute) support/support count.
- (relative) support: fraction of transactions. Frequent (> minsup)
- Confidence c: $P(Y \mid X) = \frac{\text{support}(X \cup Y)}{\text{support}(X)}$
- Subpatterns: $2^n - 1$
- Closed pattern: frequent & no super-pattern $Y \supset X$ with same support
- Max pattern: frequent & no frequent super-pattern $Y \supset X$. (更少)

-Apriori. scan k times, BFS

- Property: F itemset 子集也是F.
- Join Step: For two k-1 itemset L_1, L_2: $\{L_1[1], \dots \underbrace{L_1[k-1], L_2[k-1]}_{\text{前k-2项一样}}\}$
- Prune Step: If a candidate itemset contains infrequent subset.
- Problems: ① Multiple scan ② Huge candidates ③ Support Counting

-FPGrowth DFS, 分治.

- FP-Tree: ① Freq 1-itemset ② Sort by freq (f-list) ③ Scan, form Tree
- P35 Find Patterns having P: P-conditional Database. (Traverse FP-Tree by link)
- X-conditional FP-Tree. 注意minsup (合并 + prune). 排列组合pattern. (appendix)
- Safely remove item when its not frequent (P39)
- Benefits: complete info for frequent. No break of long pat, scan 2次.
  Reduce irrelevant info, Never larger than original DB.
  Much faster than Apriori when threshold < 1.5%.

Association Rules: For frequent pattern L, 求confidence $s \Rightarrow (L - s)$
$= \frac{\sup(L)}{\sup(s)}$

Measures

- Lift $= \frac{\sup(A \cup B)}{\sup(A) \cdot \sup(B)}$ "=1" is independent.
- Chi-Square $\chi^2 = \sum \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$, the bigger χ^2, the more related.
  Exp 是表格按sum比例分配: $e_{ij} = \frac{\text{count}(B = b_i) \times \text{count}(A = a_j)}{n}$
- All-conf$(A, B) = \min(P(A \mid B), P(B \mid A))$
- max_conf$(A, B) = \max(P(A \mid B), P(B \mid A))$
- Kulc$(A, B) = \frac{1}{2}(P(A \mid B) + P(B \mid A))$
- Cosine$(A, B) = (P(A \mid B) \cdot P(B \mid A))^{\frac{1}{2}}$

$[0, 1]$, Null-Invariant.

Sequence

- Element (event) may contain set of unordered items.
- Sequence: ordered list of events (length is events)
- Subsequence: 部分Element或Element的部分.
- Supports(α) = # of supersequence of α in DB.
- Frequent seq is "sequential pattern".

- Generalized Sequential Pattern (GSP)

- From length-1 and to len-2: ① <aa><ab><ba> ⋯ 6×6 = 36 ② <(ab)><(ac)> ⋯ 6×5/2 = 15 } S1
- ① Join S_1 & S_2: $S_1[1:] == S_2[:-1]$.

  Eg. <(12)3> + <(12)34> = <(12)34>

  <(12)3> + <2(34)> = <(12)(34)>
- ② Pruning: Check whether all k-1-supseq of a cand is in L_{k-1}
- Strength: Candidate pruning

  Drawback: Huge candidate seq, esp. 2-item.

  Multiple Scan (K)

  # of short patterns is exponential to len(mined)

  $\sum_{i=1}^{100}\binom{100}{i} = 2^{100}-1 \approx 10^{30}$ for length-100 seq pattern

- Prefix Span

- <a(ac)> is not prefix of <a(abc)(ac)d>. 字母顺序!
- α' is projection of α wrt prefix β with β

  ⇔ ① α' has prefix β ② α' is max subsequence of α
- 过程: ① Find Len-1 pat ② Form searchspace with prefix.

  ③ Mine each subset recursively in projected DB.

  注意, <(b)> 在proj DB, 不算 <ab>, 只算 <(ab)>

  同样在最后才找到prefix, 则没有suffix, 不算
- tip: ① No Candidates ② DB keep shrinking

  ③ Major cost: Construct proj DB.

DTW - Similarity Search

- Lp-norm $d(C,Q) = (\sum |C_i - q_i|^p)^{\frac{1}{p}}$. p=2 → Euclidean
- 处理Offset & scaling: Normalize, $C_i' = \frac{C_i - \mu(C)}{\sigma(C)}$
- DTW is for not line-up in X but similar shape.
- Warping Path ① 对角 ② 一个方向 ③ 单步.
- $C_p(X,Y) = \sum_l c(x_{n_l}, y_{m_l})$
- Enumerate, 3^{2N}.
- DTW: $D(n,m) = \min\{D(n-1,m), D(n,m-1), D(n-1,m-1)\} + c(x_n, y_m)$

  $D(n,1) = \sum_{k=1:n} c(x_k, y_1)$, $D(1,m) = \sum_{k=1:m} c(x_1, y_k)$

Multinomial

- W_d: sequence of words x_d: bag-of-words
- Bernoulli: $P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$

  Multinomial: $P(X_1=x_1, \dots X_k=x_k) = \frac{n!}{x_1! \cdots x_k!} \prod p_k^{x_k}$

Naive Bayes

- Bayes: $P(h|x) = \frac{P(x|h)P(h)}{P(x)}$, $P(h)$ is prior prob

  $P(x|h)$ is likelihood. $P(x)$: marginal prob. $P(h|x)$: posterior prob.
- MAP: maximum a posteriori: $h_{map} = \arg\max_{h\in H} P(h|x) = \arg\max_{h\in H} P(x|h)\cdot P(h)$.
- $P(y=j|x) = \frac{P(x|y=j)P(y=j)}{P(x)}$
- For class y, we have $\beta_y = (\beta_{y_1}, \beta_{y_2}, \dots, \beta_{y_N})$ 对应每个word分类概率.

  $P(W_{di}|y) = \beta_{yw_{di}}$, $P(W_d|y) = \prod \beta_{yw_{di}} = \prod_n \beta_{y_n}^{x_{dn}}$
- $P(y=j) = \pi_j$.
- Inference: $\theta^* = \arg\max P(D|\theta)$, $\theta \in H$, $\theta = (\beta_1, \beta_2, \dots, \beta_m, \pi)$
- $y^* = \arg\max_y p(x_d, y) = \arg\max_y p(x_d|y)p(y)$

  $= \arg\max_y \prod_n \beta_{y_n}^{x_{dn}} \cdot \pi_y$

  $= \arg\max_y \sum x_{dn} \cdot \log\beta_{y_n} + \log\pi_y$ ⇒ Linear.
- MLE, $D = \{(x_d, y_d)\}_d$, $\log L = \log \prod p(x_d, y_d|\theta) = \sum_d (x_{dn}\log\beta_{y_dn} + \log\pi_{y_d})$

  ① $\pi_j \geq 0$, $\sum \pi_j = 1$, ② $\beta_{jn} \geq 0$, $\sum_n \beta_{jn} = 1$ for $\forall j$.
- Sol: $\hat{\beta}_{jn} = \frac{\sum_{d:y_d=j} x_{dn}}{\sum_{d:y_d=j}\sum_{n'} x_{dn'}} = \frac{\text{count of word } n \text{ in class } j}{\text{count of words in class } j}$

  $\hat{\pi}_j = \frac{\sum_d 1(y_d==j)}{|D|} = \frac{\text{count of doc in class } j}{\text{count of docs}}$
- Smoothing: $\hat{\beta}_{jn} = \frac{(\sum_{d:y_d=j} x_{dn})+1}{(\sum_{d:y_d=j}\sum_{n'} x_{dn'})+N}$, still $\sum_n \hat{\beta}_{jn} = 1$.
- Eg. P21. 拉普拉斯.
- General Naive Bayes: Find y → $\max_y P(y|x) = P(y,x)/P(x) \propto P(x|y)\cdot P(y)$.

  Assumption: attributes x_i are conditionally independent for each class:

  $p(X|y) = \prod_k p(x_k|y)$. 需要Param $(3-1)\cdot 3\cdot 2+1 = 13$ (eg 13Find)
- Generative: model joint prob $p(x,y)$ ie. NB, GMM. 用参 $1+2(3^3-1)$

  Discriminative: model conditional prob $p(y|x)$. ie logistic.

Mixture Model

- Joint prob $P(x_i, z_i = C_j) = w_j f_j(x_i)$ multinomial constant → $\frac{n!}{x_1!\cdots x_k!}$
- GMM Likelihood: $L(D|\theta) = \prod_i \sum_j w_j P(x_i|\mu_j, \sigma_j^2)$, $\sum_j w_j = 1$
- Multinomial Mixture Model, $P(x_d|z=k) = \frac{(\sum_n x_{dn})!}{\prod_n x_{dn}!}\prod_n \beta_{k_n}^{x_{dn}}$
- $L = \prod_d P(x_d) = \prod_d \sum_k P(x_d, z=k) = \prod_d \sum_k P(x_d|z=k)P(z=k)$

  $= \prod_d \frac{(\sum_n x_{dn})!}{\prod_n x_{dn}!} \sum_k P(z=k) \prod_n \beta_{k_n}^{x_{dn}}$

Mixture of Unigram (seq of words)

- $P(z=k) = \pi_k$, $P(W_{di}|z=k) = \beta_{kw_{di}}$
- $L = \prod_d P(W_d) = \prod_d \sum_k P(W_d, z=k)$

  $= \prod_d \sum_k P(W_d|z=k)\cdot P(z=k)$

  $= \prod_d \sum_k P(z=k) \prod_i \beta_{kw_{di}}$, equivalent!

Probabilistic Latent Semantic Analysis (PLSA)

- β_{zw}: $P(w|z)$, $\beta_z = (\beta_{zw_1}, \beta_{zw_2}, \dots \beta_{zw_k})$, $\theta_{dz} = P(z|d)$ → soft cluster.
- Mixture of Unigrams问题: 每个document的topic唯一.

  $P(w|d) = \sum_k P(w, z=k|d) = \sum_k P(w|z=k)P(z=k|d) = \sum_k \beta_{kw}\theta_{dk}$
- Likelihood of corpus: $\prod_d P(w_1, \dots w_{N_d}, d|\theta, \beta, \pi)$

  usually uniform, $\frac{1}{M}$ $= \prod_d P(d)\left\{\prod_{n=1}^{N_d}\left(\sum_k P(z_n=k|d,\theta_d)P(w_n|\beta_k)\right)\right\}$ single word

  $= \prod_d \pi_d \left\{\prod_{n=1}^{N_d}\left(\sum_k \theta_{dk}\cdot\beta_{kw_n}\right)\right\}$

⇒ $\max \log L = \sum_{dw} c(w,d) \log \sum \theta_{dz}\beta_{zw}$, s.t. $\sum_z \theta_{dz} = 1$, $\sum_w \beta_{zw} = 1$.

E: $p(z|w,d) \propto p(w|z,d)p(z|d) = \beta_{zw}\theta_{dz}$ ⇒ $p = \frac{\beta_{zw}\theta_{dz}}{\sum_z \beta_{zw}\theta_{dz}}$

M: $\beta_{zw} \propto \sum_d P(z|w,d)c(w,d)$ 对含w的每个d求和P. eg. P26

$\theta_{dz} \propto \sum_w P(z|w,d)c(w,d)$ 对d里每个w求和P.

Linear Algebra.

z	$\frac{\partial z}{\partial x}$
Ax	A^T
x^TA	A
x^Tx	$2x$
x^TAx	$Ax + A^Tx$

$\left(\frac{u}{v}\right)' = \frac{u'v - v'u}{v^2}$, $(uv)' = u'v + v'u$

$(A^T)^{-1} = (A^{-1})^T$ $(ABC)^T = C^TB^TA^T$

$(AB)^T = B^TA^T$ $(A+B)^T = A^T + B^T$

$\det(A^T) = \det A$, $\det(A^n) = (\det A)^n$

$\det(AB) = \det A \cdot \det B$ $\det(cA) = c^n \det A$

$\partial(aX) = a \cdot \partial X$ $\partial(X+Y) = \partial X + \partial Y$

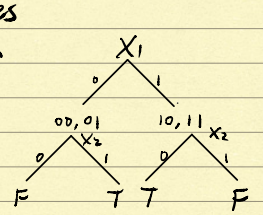
$\partial XY = (\partial X)Y + X(\partial Y)$ $(\partial X)^T = \partial X^T$

TIPS:

① Activation 是 Linear, 加入 NN still Linear

Examples

XOR



Probability.

- Empirical formula: mean − mode = 3(mean − median)
- Variance $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$

  $\sigma^2 = E[(x - E(x))^2] = E(x^2) - E(x)^2$
- z-score $= \frac{x - \mu}{\sigma}$
- Bias: $E(\hat{f}(x)) - f(x)$
- Variance: $Var(\hat{f}(x)) = E[(\hat{f}(x) - E(\hat{f}(x)))^2]$
- Covariance: $\sigma(X_1, X_2) = E(X_1 - \mu_1)(X_2 - \mu_2)$

M146 Summary

Lec5 Mistake Boundary Theorem

Perceptron: if $y(w^T x) \le 0$: $w \leftarrow w + \eta y x$

Binary, n-dimension, seperable, $\|u\| = 1$,

$\Rightarrow$ at most $(\frac{R}{\gamma})^2$ mistakes

After t mistakes: $u^T w \ge t\gamma$, $\|w\|^2 \le tR^2$,

R是图所有点的半径

① $u^T w_{t+1} = u^T w_t + y_i u^T x_i \ge u^T w_t + \gamma$.

$\Rightarrow u^T w \ge t\gamma$

② $\|w_{t+1}\|^2 = \|w_t + y_i x_i\|^2 = \|w_t\|^2 + 2y_i(w_t^T x_i) + \|x_i^2\| \le \|w_t\| + R^2$ (≤ 0)

$\Rightarrow \|w_t\| \le tR^2$

$\Rightarrow R\sqrt{t} \ge \|w_t\| \ge u^T w_t \ge t\gamma \Rightarrow t \le \frac{R^2}{\gamma^2}$ ($\|u^T\| = 1$)

Lec6 Perceptron

Aggregate the models, especially when unseparable

- Voted perceptron
  - # of vote = # of survived iterations.
  - Each iter: ①算 $y(w^T x + b)$ ② 决定update ③ 求新 w, b
  - 2 Epoches = 数据集过2遍.
- Averaged perceptron (long surviving weights also 1)
  (每一个 $a \leftarrow a + w$ 来存储和)
- Batch learning vs Online learning
  - $D = \{(x, y)\}$ — Loop for (x, y)
  - Data are i.i.d. — worst case bound
  - fast converged opt require entire

Lec7 Logistic

MLE.

Likelihood $= C_k^N \theta^k (1-\theta)^{n-k}$

$\Rightarrow \max \log(\theta^k (1-\theta)^{n-k})$

$\Rightarrow \max_\theta k \log\theta + (n-k) \cdot \log(1-\theta)$

$\frac{\partial L}{\partial \theta} = 0 \Rightarrow \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0 \Rightarrow \theta(n-k) = (1-\theta)k$

$\Rightarrow \theta = \frac{k}{n}$

MLE of logistic Reg:

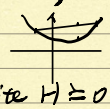
$P(y|w, x) = \sigma(y_i w^T x_i)$

$\Rightarrow \max_w \sum^N -\log(1 - \exp(-y_i w^T x_i))$

$P(S|w) = \arg\max_w \prod P(y=1|x_i, w)^{y_i} P(y=0|x, w)^{1-y_i}$

$\Rightarrow \arg\max_w \sum y_i \log \sigma(w^T x) + (1-y_i)\log(1-\sigma(w^T x))$

Convex: ① $f(x)$ is convex if $f''(x) \ge 0$



② $f(\lambda a + (1-\lambda)b) \le \lambda f(a) + (1-\lambda) f(b)$.

③ If Hessian is positive semi-definite $H \succeq 0$

i.e. $z^T H z \ge 0$ for $\forall z$.

- Gradient Decent $\theta^{t+1} \leftarrow \theta^t - \eta \nabla J(\theta^t)$
- Stochastic: By a gradient at a single example at a time

Lec10 Probably Approximately Correct (PAC)

$m > \frac{n}{\varepsilon}(\log(n) + \log\frac{1}{\delta})$, for m iid examples

instance of length n, error $\le \varepsilon$, prob $\ge (1-\delta)$

$P(\text{bad literal survive 1 example}) < 1 - \frac{\varepsilon}{n}$.

$P(\text{Any bad literal survive m exps}) < n(1-\frac{\varepsilon}{n})^m < \delta$

Since $1 - x < e^{-x} \Rightarrow ne^{-\frac{m\varepsilon}{n}} < \delta$

$\Rightarrow \frac{m\varepsilon}{n} > -\log\frac{\delta}{n} \Rightarrow m > \frac{n}{\varepsilon}(\log n + \log\frac{1}{\delta})$

- Real Expectation: With high prob it learn a close approximation to the target concept
- Hypothesis is Consistent with a training set if:

$$m > \frac{1}{\varepsilon}(\ln|H| + \ln\frac{1}{\delta})$$

Lec 11.

- Hypothesis space: 2^{2^n} possible functions

If Conjunctive, $C_4^0 + C_4^1 + C_4^2 + C_4^3 + C_4^4 = 16$ possible functions

- Union bound: Prob of at lease 1 of events happen < the sum of probs of the events

$\Rightarrow err_D(h) \le \sum_{z \in h} p(z)$

- Shatter: S of examples is shattered by set of functions H if for every partition of exps in S into pos & neg exps, there's a func in H gives exact label.
- Vapnik-Chervonenkis Dimension

  VC-D of H over X is the size of largest finite subset of X that shattered by H.
- Eg. 线, 4个点不行. VC = 3.

Lec 12 Multi Class

One-against-All, $O(K)$

- $y = \arg\max(W_a^T x, W_b^T x, W_c^T x)$
- Not always possible to learn. Assumption: Each class separable
- No theo Justification: Range of all classifier need be same
- Easy implement, well in practice. ^ set unbalanced

One-vs-One, $O(K^2)$

- $C(K, 2)$ binary classification
- Each label gets $K-1$ votes

Majority: x with label i if i wins more often than $j = 1 \sim K$.

Tournament: start with n/2 pairs. (like世界杯)

- Every pair of classes is separable.
- Overfitting, large space for model

Logistic Loss: $\log(1 + \exp(-y_i w^T x_i))$

Regulation:

L_0: $R(w)$ = # non-zero element in w

L_1: $R(w) = \sum_i |w_i|$

L_2 (Gaussian prior) $R(w) = \sum_i w_i^2 = w^T w$