k-means

1. Given sample $\mathcal{X} = \{x^t\}_{t=1}^N$
   Find k reference vectors m_j which best represent $\mathcal{X}$.

Encoding / Decoding

$i = \arg\min_j \| x^t - m_j \|$.

label : $b_i^t = \begin{cases} 1. & i = \arg\min_j \| x^t - m_j \| \\ 0. & \text{otherwise} \end{cases}$

reconstruction error: $E(\{m_j\}_{j=1}^k \mid \mathcal{X}) = \sum_t \sum_i b_i^t \| x^t - m_i \|^2$

2. Optimization.

$$\underset{\{m_i\}_{i=1}^k, \{b^t\}_{t=1}^N}{\text{minimize}} \sum_t \sum_i b_i^t \| x^t - m_i \|^2.$$

subject to. $b_i^t = \begin{cases} 1. & i = \arg\min_j \| x^t - m_j \| \\ 0. & \text{otherwise} \end{cases}$

* b_i^t depends on m_i. no analytical. but iterative

3. Algorithm.

Initialize. $\{m_i\}_{i=1}^k$. (e.g. random ~~x~~ x^t).

Repeat.

For all $x^t \in \mathcal{X}$. obtain estimated label b^t.

For all m_i. $i = 1 .. k$. (take derivative, and $=0$)

$$m_i = \frac{\sum_t b_i^t x^t}{\sum b_i^t}$$ 更新 reference. m_j.

Until converge

4. Remark: converge in finite iters.
   final m_i highly depends on init m_i

Overview

1. Least square. $RSS = \| y - X\beta \|_2^2$.

$y \in R^N$ $X \in R^{N \times p}$. $\beta \in R^p$

$$\frac{\partial RSS}{\partial \beta} = 2X^T y - 2X^T X \beta = 0 \Rightarrow \hat{\beta} = (X^T X)^{-1} X^T y$$

2. Nearest neighbour.

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x_0)} y_i$$

~~3. Statistical Decision theory~~

~~Expected prediction error (EPE)~~

$\cancel{EPE(f) = E(Y - f(X))^2}$

$\cancel{= \iint (y - f(x))^2 f_{X,Y}(x, y)\, dx\, dy}$

$\cancel{f_{X,Y}(x,y) = f_{Y|X}(y|x) f_X(x)}$ ~~Adam's Law~~

$\cancel{\Rightarrow EPE(f) = E_X(E_{Y|X}((Y - f(x))^2 \mid X))}$

~~minimize EPE pointwise:~~

$\cancel{f(x) = \arg\min_c E_{Y|X}((Y-c)^2 \mid X = x)}$

~~regression function.~~ $\cancel{f(x) = E(Y \mid X = x)}$

3. Statistical decision theory.

$(X, Y) \sim Pr(X, Y)$.

$f(X) \Rightarrow Y$

$\min_f EPE(f)$

Regression: $\min_f E[L(Y, f(X))]$

- L_2: $\hat{f}(x) = E(Y \mid X = x)$
  - Parametric: $\hat{\beta} = (X^T X)^{-1} X^T y$
  - Non ...: $\hat{f}(x) = Ave(y_i \mid x_i \in N_k(x))$
- L_1: $\hat{f}(x) = \text{median}(Y \mid X = x)$

Classification: $\min_f E[L(G, \hat{G}(X))]$

- zero-one loss: $\hat{G}(x) = \arg\max_{k \in \mathcal{G}} Pr(G = k \mid X$

4. Local Methods in High dimensions

bias-variance decomposition

1. Deterministic. * $f(x_0)$: g.t, $\hat{y}_0$: pred value

~~MSE(x₀) =~~

$EPE(x_0) = MSE(x_0)$

$= E(f(x_0) - \hat{y}_0)^2$.

$= E(f(x_0) - E_T(\hat{y}_0) + E_T(\hat{y}_0) - \hat{y}_0)^2$

$= E(f(x_0) - E_T(\hat{y}_0))^2 + E(E_T(\hat{y}_0) - \hat{y}_0)^2 \rightarrow Var$ (Bias)

$+ 2E((f(x_0) - E_T(\hat{y}_0))(E_T(\hat{y}_0) - y_0)) \rightarrow 0$.

$= Var_T(\hat{y}_0) + Bias^2(\hat{y}_0)$

2. Non-Deterministic

$EPE(x_0) = MSE(x_0) + \sigma^2$

$= Var(\hat{y}_0) + Bias^2(\hat{y}_0) + \sigma^2$

Linear regression.

1. ridge regression.

$$\hat{\beta}^{ridge} = \arg\min_\beta \left\{ \sum_i (y_i - \beta_0 - \sum_j x_{ij}\beta_j)^2 + \lambda \|\beta\|_2^2 \right\}$$

or. $$\hat{\beta}^{ridge} = \arg\min_\beta \left\{ \sum_i (y_i - \beta_0 - \sum_j x_{ij}\beta_j)^2 \right\}.$$

subject to $\|\beta\|_2^2 \le t$

closed form. $\hat{\beta}^{ridge} = (X^T X + \lambda I_p)^{-1} X^T y$

SVD: $U \in R^{N \times p}$ (column space of X) $X = UDV^T$

$V \in R^{p \times p}$ (*row space of X) $U^T U = I.\ V^T V = I$

$D \in R^{p \times p}$ (diagonal. singular values)

Least square:

$$X\hat{\beta}^{LS} = X(X^T X)^{-1} X^T y = UU^T y = \sum_j u_j u_j^T y$$ (u_j: j-th column of U.)

Ridge.

$$X\hat{\beta}^{ridge} = X(X^T X + \lambda I)^{-1} X^T y$$

$$= UD(D^2 + \lambda I)^{-1} D U^T y$$

$$= \sum_j u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T y$$

2. Lasso

岭回归. $L_2 \Rightarrow L_1$ 正则项

Linear Classification.

~~1. Linear Discriminant Analysis~~

$$Pr(G=k|X=x) = \frac{Pr(X=x|G=k)Pr(G=k)}{Pr(X=x)}$$

Density. X in $G=k$: $f_k(x) = Pr(X=x|G=k)$.

Prior. $\pi_k = Pr(G=k)$

1. Linear Discriminant Analysis.

Model density as MVN.

$$\hat{f}_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right)$$

Assume each class share a common covariance $\Sigma_k = \Sigma$

Compare class k & l. $=0 \Rightarrow$ Decision boundary

$$\log\frac{Pr(G=k|X=x)}{Pr(G=l|X=x)} = \log\frac{f_k(x)}{f_l(x)} + \log\frac{\pi_k}{\pi_l}$$

$$= \log\frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k+\mu_l)^T\Sigma^{-1}(\mu_k-\mu_l) + x^T\Sigma^{-1}(\mu_k-\mu_l)$$

Parameter estimation: $\hat{\pi}_k = \frac{N_k}{N}$. $\hat{\mu}_k = \sum_{g_i=k}\frac{x_i}{N_k}$.

$$\hat{\Sigma} = \sum_{k=1}^{K}\sum_{g_i=k}(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^T/(N-K)$$

$$= \frac{(N_1-1)\hat{\Sigma}_1 + \cdots + (N_k-1)\hat{\Sigma}_k}{(N_1-1)+\cdots+(N_k-1)}$$

$$\delta_k(x) \triangleq x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k$$

2. Quadratic ~~Linear~~ Discriminant Analysis

$$\delta_k(x) \triangleq -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log\pi_k$$

* each class specific covariance Σ_k

· Fisher's Formula (LDA)

Eigen decomposition. $\hat{\Sigma} = UDU^T$.

$$\delta_k(x) \propto Pr(G=k|X=x) = -\frac{1}{2}(x-\hat{\mu}_k)\hat{\Sigma}^{-1}(x-\hat{\mu}_k) + \log\hat{\pi}_k + C.$$

$$= -\frac{1}{2}\|x^* - \hat{\mu}_k^*\|_2^2 + \ln\hat{\pi}_k + C.$$

$$x^* = D^{-\frac{1}{2}}U^Tx.\quad \hat{\mu}_k^* = D^{-\frac{1}{2}}U^T\hat{\mu}_k$$

4. Logistic regression.

Model: $\log\frac{Pr(G=l|X=x)}{Pr(G=K|X=x)} = \beta_{l0} + x^T\beta_l$

$$\Rightarrow Pr(G=l|X=x) = \frac{\exp(\beta_{l0}+x^T\beta_l)}{1+\sum_{i=1}^{K-1}\exp(\beta_{i0}+x^T\beta_i)}$$

$$Pr(G=K|X=x) = \frac{1}{1+\sum_{i=1}^{K-1}\exp(\beta_{i0}+x^T\beta_i)}$$

Parameter set $\theta = \{\beta_{10}, \beta_1 \cdots \beta_{(K-1)0}, \beta_{K-1}\}$

MLE. $\ell(\theta) = \log Pr(\vec{g}|X;\theta) = \sum_{i=1}^{N}\log Pr(\overset{\text{label}}{g_i}|x_i;\theta)$

Probability and Estimation

1. Naive Bayes

$$P(X_1,\ldots,X_n|Y) = \prod_i P(X_i|Y)$$

assume X_i are conditionally indep given Y.

2. Naive Bayes Algorithm. discrete X_i

Train:

for each y_k

estimate $\pi_k = P(y=y_k)$

for each $x_{ij} \in X_i$.

estimate $\theta_{ijk} \equiv P(X_i=x_{ij}|Y=y_k)$

Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} P(Y=y_k)\prod_i P(X_i^{new}|Y=y_k)$$

$$= \arg\max_{y_k} \pi_k \prod_i \theta_{ijk}$$

3. Estimate parameters. MLE.

$$\hat{\pi}_k = \frac{\#D\{Y=y_k\}}{|D|}$$

$$\theta_{ijk} = \frac{\#D\{X_i=x_{ij}, Y=y_k\}}{\#D\{Y=y_k\}}$$

4. Estimate paras. MAP.

data not in D. $\Rightarrow$ MLE estimate $P(X_i|Y)=0$

$$\hat{\pi}_k = \frac{\#D\{Y=y_k\}+(\beta_k-1)}{|D|+\sum_m(\beta_m-1)}$$

$$\hat{\theta}_{ijk} = \frac{\#D\{X_i=x_{ij}, Y=y_k\}+(\beta_k-1)}{\#D\{Y=y_k\}+\sum_m(\beta_m-1)}.$$

5. Continuous (Gauss. Naive Bayes)

assume.

$$P(X_i=x|Y=y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{(x-\mu_{ik})^2}{\sigma_{ik}^2}\right)$$

Algorithm similar to discrete case

6. Estimate Paras

$$\hat{\mu}_{ik} = \frac{\sum_j X_i^j\,\delta(Y^j=y_k)}{\sum_j \delta(Y^j=y_k)}.$$

j-th sample

i-th feature

k-th class.

$\delta(z)=1$ if z =

$$\sigma_{ik}^2 = \frac{1}{\sum_j\delta(Y^j=y_k)}\sum_j(X_i^j-\hat{\mu}_{ik})^2\delta(Y^j=y_k)$$

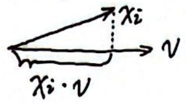
PCA

1. $v_1 \dots v_d$: d principle ~~vectors~~ components. $v_i \cdot v_i = 1$. $v_i \cdot v_j = 0$.

$X = [x_1 \dots x_n]$. data (centered)

maximize sample variance: $\frac{1}{n}\sum (v^T x_i)^2 = v^T X X^T v$

maximize $v^T X X^T v$. s.t. $v^T v = 1$.



Lagrangian: $\max\limits_{v}.\ v^T X X^T v - \lambda v^T v$

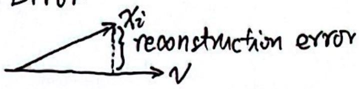
$\frac{\partial}{\partial v} = 0 \Rightarrow (XX^T - \lambda I) v = 0$. ie. $(XX^T) v = \lambda v$.

2. v: eigen vector of sample corr/cov matrix XX^T.

Sample variance of projection $v^T X X^T v = v^T \lambda v = \lambda$

3. Minimum Reconstruction Error

$\frac{1}{n}\sum_{i=1}^{n} \| x_i - (v^T x_i) v \|^2$



勾股定理

4. 0 eigenvalue $\Rightarrow$ no variability along those direction.

Only keep data projections onto non-zero eigenvalue components. $v_1 \dots v_k$. $k = \mathrm{rank}(XX^T)$

$x_i = (x_i^1, \dots x_i^D) \Rightarrow (v_1 \cdot x_i, \dots v_d \cdot x_i)$

Only keep data projections onto large eigenvalue $\Leftrightarrow$ ignore components of small signifigance (noise)

Gradient Descent

1. ~~probel~~ problem: $\min\limits_{x \in R^n} f(x)$.

Iteration: $x^{r+1} = x^r - \gamma_r \cdot \nabla f(x^r)$

2. Convex: $f(\lambda x + (1-\lambda) y) \le \lambda f(x) + (1-\lambda) f(y)$

$f(x) \ge f(y) + \nabla f(y)^T (x-y)$

$\nabla^2 f(x) \succeq 0$.

3. L-smooth: $\|\nabla f(x) - \nabla f(y)\| \le L \|x-y\|$

4. Descendent Lemma. $|f(x) - f(y) - \nabla f(y)^T (x-y)| \le \frac{L}{2}\|x-y\|^2$

f: twice differentiable. L-smooth $\Leftrightarrow \nabla^2 f(x) \preceq LI$. $d^T \nabla^2 f(x) d \le L\|d\|^2$. $\forall d$.

5. Convergence analysis:

Optimality measure: $M(x^r)$ $\begin{cases} \text{convex: } \|x^r - x^*\|.\ f(x^r) - f^*. \\ \text{non-convex: } \|\nabla f(x^r)\|. \end{cases}$

Order of convergence q. s.t. $\sup\{q \mid \lim\limits_{r\to+\infty} \frac{M(x^{r+1})}{M(x^r)^q} < \infty\}$

$q=1$: linear convergence. $q=2$: quadratic.

Rate of convergence: given q. $\lim\limits_{r\to\infty} \frac{M(x^{r+1})}{M(x^r)^q} = \eta$

Sublinear: $\lim\limits_{r\to\infty} \frac{M(x^{r+1})}{M(x^r)} = 1$. Superlinear: $\lim\limits_{r\to\infty} \frac{M(x^{r+1})}{M(x^r)} = 0$

4. Convergence under convexity.

Polyak's stepsize $\gamma_r = \frac{f(x^r) - f^*}{\|\nabla f(x^r)\|^2}$

$\|x^{r+1} - x^*\|^2 \le \|x^r - x^*\|^2 - \frac{(f(x^r) - f^*)^2}{\|\nabla f(x^r)\|^2}$

Th. f: convex. $\|\nabla f\| \le B$. $(x^r)_{r\in N}$. generated by polyaks step size satisfies

$\min\limits_{r=0,\dots T-1} f(x^r) - f^* \le \frac{B\|x^0 - x^*\|}{\sqrt{T}}$.

Fix r, optimal $\gamma^* = \frac{\|x_0 - x^*\|}{\sqrt{T} B}$

5. Convergence under smoothness

convex upper bound (quadratic)

$f(x) \le f(y) + \nabla f(y)^T (x-y) + \frac{L}{2}\|x-y\|^2$

minimize by $\gamma = \frac{1}{L}$.

$x^{r+1} = x^r - \gamma \nabla f(x^r)$

$= \arg\min\limits_{x} \underbrace{\{f(x^r) + \nabla f(x^r)^T (x - x^r) + \frac{1}{2\gamma}\|x - x^r\|^2\}}_{L(x \mid x^r)}$

$\gamma \le \frac{1}{L}$: $L(x \mid x^r) \ge f(x)$

By descendent Lemma, $\gamma \le \frac{1}{L}$. $x^{r+1} = x^r - \gamma \nabla f(x^r)$

$\Rightarrow f(x^{r+1}) \le f(x^r) - \frac{\gamma}{2}\|\nabla f(x^r)\|^2$

$\gamma < \frac{2}{L}$: $f(x^{r+1}) \le f(x^r) - \gamma(1 - \frac{\gamma L}{2})\|\nabla f(x^r)\|^2$

Th. f: L-smooth. $\gamma \le \frac{1}{L}$.

$\min\limits_{r=0,\dots,T-1} \|\nabla f(x^r)\|^2 \le \frac{\frac{2}{\gamma}(f(x^0) - f(x^*))}{T}$

6. Convexity & smoothnees.

$\|x^{r+1} - x^*\|_2^2 = \|x^r - \gamma \nabla f(x^r) - x^*\|^2$

$= \|x^r - x^*\|^2 - 2\gamma \underbrace{\nabla f(x^r)^T (x^r - x^*)}_{\text{convexity}} + \underbrace{\gamma^2 \|\nabla f(x^r)\|}_{\text{smooth}}$

$\le \|x^r + x^*\|^2 - 2\gamma (f(x^{r+1}) - f^*)$

Strong Convexity (μ)

$f(\lambda x + (1-\lambda) y) \le \lambda f(x) + (1-\lambda) f(y) - \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|^2$

$f(x) \ge f(y) + \nabla f(y)^T (x-y) + \frac{\mu}{2}\|x-y\|_2^2$

$\nabla^2 f(x) \succeq \lambda I$

Upper & lower bound.

$f(x) \ge f(y) + \nabla f(y)^T (x-y) + \frac{\mu}{2}\|x-y\|_2^2$

$f(x) \le f(y) + \nabla f(y)^T (x-y) + \frac{L}{2}\|x-y\|^2$

implication: $\begin{cases} \nabla f(x^r)^T (x^r - x^*) \ge f(x^r) - f^* + \frac{\mu}{2}\|x^r - x^*\|^2 \\ f(x^{r+1}) \le f(x^r) - \frac{\gamma}{2}\|\nabla f(x^r)\|^2. \end{cases}$

Lagrangian.

1. minimize $f_0(x)$. (optimal $\cancel{f_0}$ p^*)
   subject to $f_i(x) \le 0$. $i=1,\dots,m$.

   $\mathcal{L}(x,\lambda) = f_0(x) + \lambda_1 f_1(x) + \dots + \lambda_m f_m(x)$.

2. dual function
   $g(\lambda) = \inf_x \mathcal{L}(x,\lambda)$.

   lower bound property: if $\lambda \ge 0$. x: primal feasible
   $g(\lambda) \le f_0(x)$

3. dual problem.
   maximize $g(\lambda)$. (optimal d^*)
   subject to $\lambda \ge 0$.
   $d^* \le p^*$, $p^* - d^*$: ~~dual gap~~ optimal dual gap.
   convex problem $\Rightarrow p^* = d^*$

4. KKT ~~o~~Optimal Condition.
   $f_i(x^*) \le 0$. (primal feasible)
   $\lambda_i^* \ge 0$. (dual feasible)
   $\lambda_i^* f_i(x^*) = 0$. (complementary)
   $\nabla f_0(x^*) + \sum \lambda_i^* \nabla f_i(x^*) = 0$ (stationary)

5. Equality constraints
   minimize $f_0(x)$
   subject to. $f_i(x) \le 0$. $i=1,\dots,m$
   $h_i(x) = 0$. $i=1,\dots,p$

   $\mathcal{L}(x,\lambda,\nu) = f_0(x) + \sum \lambda_i f_i(x) + \sum \nu_i h_i(x)$

   dual function: $g(\lambda,\nu) = \inf_x (\mathcal{L}(x,\lambda,\nu))$

   dual problem: maximize $g(\lambda, \nu)$.
   subject to $\lambda \ge 0$

   KKT. $f_i(x^*) \le 0$. $h_i(x^*) = 0$.
   $\lambda_i^* \ge 0$.
   $\lambda_i^* f_i(x^*) = 0$.
   $\nabla f_0(x^*) + \sum \lambda_i^* \nabla f_i(x^*) + \sum \nu_i^* \nabla h_i(x^*) = 0$