Recap Exp design

Chosing $\underbrace{x_1 \ldots x_k}_{S}$ out of $n$ s.t. $R(\hat{f}_S) - R(f^*)$ $\leftarrow$ $Y \approx f^*(x) + \varepsilon$, $f^*(x) = \theta^{*T} x$

$\varepsilon \sim$ iid $\mathcal{N}(0, \sigma^2)$

linear regression $= E[\langle \hat{\theta}_S - \theta^*, x \rangle]$

$= x^T (X_S^T X_S)^{-1} x$ $\Leftarrow$ predictor var

A-opt $\quad E[(\hat{\theta}_S - \theta)^2] \equiv (X_S^T X_S)^{-1}$ var of parameters

E-opt $\quad \max_x x^T (X_S^T X_S)^{-1} x$

V-opt $\quad X^T (X_S^T X_S)^{-1} X$

$$f((X_S^T X_S)^{-1}) = f((X^T W X)^{-1})$$

combinatorial opt $\quad W = \begin{bmatrix} 0 & & 0 \\ & 1 & \\ & & 0 \\ 0 & & & 1 & \\ & & & & \ddots \end{bmatrix} = \mathrm{diag}(w)$

$w_i \in \{0,1\}$

$\sum w_i = k$

Continuous relaxation

$w_i \in [0,1]$ $\leftarrow$

$\exists$ a vector $w'$ that is $O(d^2)$ sparse s.t. $X^T W X = X^T W' X$

$\sum w_i' = 1 \quad \sum w_i = 1$

Optimal exp design procedure G-opt

1. Solve cont relaxed version of G-opt to get $w$.
2. Has $O(d^2)$ sparse $w'$
3. Sample $i^{th}$ data points $n_i = \lceil w_i' k \rceil$ times $(x_i, y_i)$
4. Build est $\hat{\theta}$ using these pts.

Thm: Wp $\geq 1 - \delta$ $\quad \forall x \in \mathcal{X}$ $\quad |\langle x, \hat{\theta} - \theta^* \rangle| \leq \sqrt{\|x (X_S^T X_S)^{-1} x\|} \sqrt{2\sigma^2 \log(2|\mathcal{X}|/\delta)}$

$= O(d/k)$ $\quad \simeq \sqrt{\frac{d}{k}}$

where $\hat{\theta}$ constructed using

$\ldots \lceil w_i' k \rceil \ldots \sum \ldots 1 + w_i' k = O(d^2) + k$

$$\sum_i n_i = \sum_{i \sim O(d^2)} \lceil w_i' k \rceil \le \sum_{i \sim O(d^2)} \ldots$$

$(\because \sum w_i' = 1)$

data points/labels.

$$\star \quad x^T (X_S^T X_S)^{-1} x = x^T \Big( \sum_i n_i x_i x_i^T \Big)^{-1} x \le x^T \Big( \sum_{i \sim O(d^2)} \lceil w_i' k \rceil x_i x_i^T \Big)^{-1} x$$

$$\le x^T \underbrace{\Big( \sum_i w_i' x_i x_i^T \Big)^{-1} x}_{k} \leftarrow O(d)$$

$$\max_x x^T \Big( \sum_i w_i' x_i x_i^T \Big)^{-1} x \ge \sum_i w_i' x_i^T (X^T W' X)^{-1} x_i = \mathrm{tr}\big(X^T W' X (X^T W' X)^{-1}\big)$$

$\max \ge$ ~~sum~~ avg

$$= \mathrm{tr}(I) = d$$

Since $w'$ is optimal, achieves lower bound with equality $\supseteq$
(Kiefer-Wolfowitz Equivalence Theorem)

---

Better rounding techniques

regret min

$1+\varepsilon$ using $k = \Omega\left(\frac{d^2}{\varepsilon}\right)$ or $\Omega\left(\frac{d}{\varepsilon^2}\right)$ $\forall$ A, D, T, E, V, G

$1+\varepsilon$ using $k = \Omega\left(\frac{d}{\varepsilon}\right)$ only A, D opt. $\leftarrow$ Federov's algo.
Greedy algo.

Federov - randomly sampling $k$ pts. &
swapping pts.

Greedy - small randomly samples.
greedy addition.

---

Nonlinear regression

- Generalized linear regression $\quad g(f^*(x)) = \theta^{*T} x \quad$ $g$ - form known

$Y = f^*(x) + \varepsilon$

$$f^*(x) = \frac{e^{\theta^{*T} x}}{1 + e^{\theta^{*T} x}}$$

var of prediction $\hat{\theta}$ - MLE

under mild reg. MLE $\quad E\left[\|\hat{\theta} - \theta^*\|^2\right] = (1 + o(1))\, \mathrm{tr}\big(I(X, \theta^*)^{-1}\big)$

Fisher information matrix

linear $I(X,\theta^*) = X^T X$

$$\min_{|S| \le k} tr(I(X_S,\theta^*))^{-1}$$

1. k/2 samples randomly → get estimate $\hat{\theta}_{MLE} \leftarrow k/2$

2. plug-in

$$\tilde{x}_i \leftarrow func(x_i, \hat{\theta}, g)$$

- Neural Networks (deep) Core-set sampling.

$$\left| \frac{1}{n}\sum_{i=1}^{n} l(x_i,y_i) - \frac{1}{|S|}\sum_{i\in S} l(x_i,y_i) \right| \le \text{small}$$

(don't want to use $\{y_i\}_{i=1}^n$)

loss $l$ is Lipschitz

≡ K-center problem

$$\min_{|S'|\le k} \max_{i} \min_{j\in S'\cup S^0} d(x_i,x_j)$$

$S^0$ - initial ~~labeled~~ dataset

$S'$ - new labeled dataset

NP-hard

greedy solⁿ - approxⁿ ratio 2

---

Active Learning - sequentially choose $x_1 \ldots x_T$ to min $R(\hat{f}_{x_1 \ldots x_T}) - R(f^*)$ ★

Regression $x_t = \arg\min_x \hat{\sigma}_{t-1}(x)$

↳ closed form linear models

generalized linear models

Bayesian models (GP)

ensembles for general nonlinear models

e.g. NNs.

★ $\min_{x_t} R(\hat{f}_{x_1 \ldots x_t, y_1 \ldots y_{t-1}})$ ~~$-R(f^*)$~~

$\hat{f}_1 \ldots \hat{f}_m$

↓

$x \rightarrow \hat{y}_1 \quad \hat{y}_m$

$$\underset{=}{x} \to \frac{1}{m}\sum_{j=1}^{m}\Big(y_j - \frac{1}{m}\sum_{j=1}^{m} y_i\Big)^2$$

Classification

uncertainty of predicted labels

Binary classes $\quad Ber(p) \sim p(1-p) \quad$ max $p = \frac{1}{2}$

$$x \to \underset{\sim}{P(Y=1|x)},\ \underset{\sim}{P(Y=0|x)}$$

Logistic regression $\quad P(Y=1|X) = \dfrac{e^{\theta^T x}}{1+e^{\theta^T x}}$

$P(Y=1|X) - \frac{1}{2}$ ← near decision boundary

Multiple classes $\quad$ least confident

$$\arg\min_x \; 1 - \max_y P(y|x)$$

margin sampling

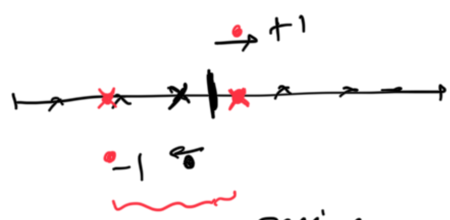
$$\arg\min_x P(Y_{(1)}|x) - P(Y_{(2)}|x)$$

entropy sampling

$$\arg\max_x \sum_y P(Y=y|x)\log\frac{1}{P(Y=y|x)}$$

How to extend these ideas to non-probabilistic classifiers?

Linear, SVM, Decision Trees




1-dim $x$

no noise $\quad$ passive $\quad 1/n$

active $\quad 1/2^n$

d-dim linear

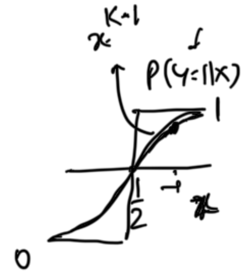
no noise $\begin{cases} \text{passive} & d/n \\ \text{active} & e^{-n/d} \end{cases}$

with noise $\begin{cases} \text{passive} & \left(\frac{d}{n}\right)^{\frac{k}{2k-1}} \\ \text{active} & \left(\frac{d}{n}\right)^{\frac{k}{2k-2}} \end{cases}$

$\rightarrow$

Algo adaptive noise

linear case.




Tsybakov noise label