427- Lecture 11

- Naive Bayes & Feature Selection
- ℓ_1 - Regularization $\leftrightarrow$ LASSO $\leftrightarrow$ Compressed Sensing.
  - Application Well

Last Time: Model Selection using BIC

$$P(\theta \mid \mathcal{D}, \mathcal{F}) \overset{\text{Bayes Rule}}{=} \frac{L(\mathcal{D}; \theta, \mathcal{F})\, P(\theta \mid \mathcal{F})}{P(\mathcal{D} \mid \mathcal{F})} = \frac{\cdots\cdots}{\int_\theta L(\mathcal{D}; \theta, \mathcal{F})\, P(\theta \mid \mathcal{F})\, d\theta}$$

- θ: Candidate
- $\mathcal{F}$: Func. class
- $L(\mathcal{D}; \theta, \mathcal{F})$: likelihood
- $P(\theta \mid \mathcal{F})$: prior
- $P(\theta \mid \mathcal{D}, \mathcal{F})$: linear, quad., R.B.F
  - posterior
- $P(\mathcal{D} \mid \mathcal{F})$: Model Evidence (Bayesian Score)
  - Not BIC $\leftarrow$ Just an Approximation

Seek $\rightarrow$ maximize $P(\mathcal{D} \mid \mathcal{F}_i)$ over $f \hat{=}$ class $\mathcal{F}_i$ $i = 1, 2, \ldots, K$.

$\Downarrow$

Intractable!

$\uparrow$

Approximate Log $P(\mathcal{D} \mid \mathcal{F}_i)$ by $BIC(i)$

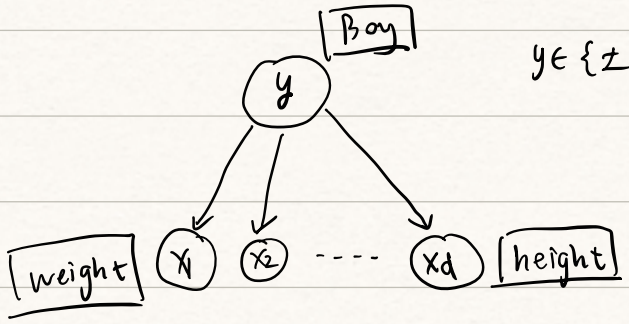
$$BIC(i) = \log L(\mathcal{D}; \hat{\theta}, \mathcal{F}_i) - \frac{d_{\mathcal{F}_i}}{2} \log n$$

- $d_{\mathcal{F}_i}$: # of parameters in model $\mathcal{F}_i$
- $\log L(\mathcal{D}; \hat{\theta}, \mathcal{F}_i)$: maximize likelihood
- $\frac{d_{\mathcal{F}_i}}{2} \log n$: penalty

$\Rightarrow$ Pick Model $i \in \{1, \ldots, K\}$ if $i = \arg\max_{i=1,\ldots,K} BIC(i)$

This time: Feature selection ← { Naive Bayes
ℓ_1 − reg.

Naive Bayes → Graphic Model



$y \in \{\pm 1\}$

AIM:
Get smaller model !

Remove the useless feature

Suppose features in each trg. sample $\underline{X} = (X_1, \ldots, X_d) \in \{\pm 1\}^d$ are related to the label $y \in \{\pm 1\}$ as follows:

(Naive Bayes) Assumption

$$P(\underline{X}, y) = P(y)\, P(\underline{X} \mid y) = P(y) \prod_{i=1}^{d} P(X_i \mid y)$$

i.e. the features $X_1, \ldots, X_d$ are conditionally independent Given y.

[Example]: if you are a Boy, your height and weight are independent

Compare the NB model to an even simpler 'NULL' model

Assp.: Feature all independent of y. (label)

$$P(\underline{X}, y) = P(y)\, P(\underline{X} \mid y) = P(y)\, P(\underline{X})$$
$$= P(y) \prod P(X_i)$$

Assuming NB Model, maximize the log-likelihood of $\mathcal{D}_n$ Given model

↓

① $\mathcal{D}_n = \{(\underline{x}_t, y_t)\}_{t=1}^{n} \in \{\pm 1\}^d \times \{\pm 1\}$

② $\ell(\mathcal{D}_n; \theta) = \sum_{t=1}^{n} \log P(\underline{x}_t, y_t) \longrightarrow$ [Independence across $(\underline{x}_t, y_t)$]

$$\text{(NB Asumption)} = \sum_{t=1}^{n} \log\left[P(y_t) \prod_{i=1}^{d} P(x_{ti} \mid y_t)\right]$$

$$= \sum_{t=1}^{n} \left[\log P(y_t) + \sum_{i=1}^{d} \log P(x_{ti} \mid y_t)\right]$$

$$= \sum_{i=1}^{d} \sum_{t=1}^{n} \log P(\underbrace{x_{ti}}_{\pm 1} \mid \underbrace{y_t}_{\pm 1}) + \sum_{t=1}^{n} \log P(y_t) \qquad (*)$$

[Trick]

$$\sum_{t=1}^{n} \log P(y_t) = \underbrace{n_y(1)}_{\#\text{ of } y_t=1} \log P(1) + \underbrace{n_y(-1)}_{\#\text{ of } y_t=-1} \log P(-1)$$

$$\Downarrow$$

Since $y_t \in \{+1, -1\}$ $\quad (n_y(1) + n_y(-1) = n)$

$$\sum_{y \in \{-1,+1\}} n_y(y)\, P(y)$$

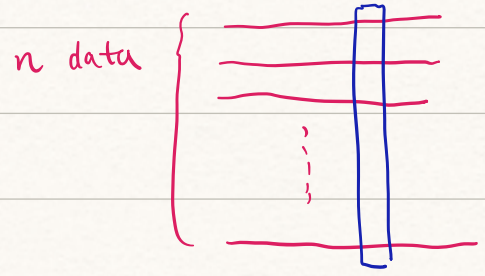
Thus, $(*)$ can be rewritten as:

$\ell(\mathcal{D}_n; \theta)$ → $\frac{P(x_i \mid y)}{P(y)}$

$$\sum_{i=1}^{d} \sum_{(x_i, y) \in \{\pm 1\} \times \{\pm 1\}} \hat{n}_{iy}(x_i, y) \log P(x_i | y) + \sum_{y \in \{\pm 1\}} \hat{n}_y(y) \log P(y)$$

A: Meaning d feature

B: parameter θ



from data

$x_{1i}, \ldots\ldots x_{ni}$

$x_1, \ldots\ldots, x_n$

$\hat{n}_{iy}(x_i, y)$ ⇒ # of $x_i = \pm 1$, $y = \pm 1$ for all n data

$\hat{n}_y(y)$ ⇒ # of $y = \pm 1$ for all n data

→ Differentiate $\ell(\mathcal{D}_n; \theta)$ with $P(x_i|y)$ & $P(y)$ to get MLE of these parameters.

PART B: $B = \hat{n}_y(-1) \log P(-1) + \hat{n}_y(1) \log P(1)$

$= \hat{n}_y(-1) \log (1 - P(1)) + \hat{n}_y(1) \log P(1)$

$$\frac{\partial B}{\partial P(1)} = \hat{n}_y(-1) \frac{-1}{1 - P(1)} + \hat{n}_y(1) \frac{1}{P(1)} = 0$$

$$\Leftrightarrow -(n - \hat{n}_y(1)) \frac{1}{1 - P(1)} + \hat{n}_y(1) \frac{1}{P(1)} = 0$$

$$\Leftrightarrow -P(1)(n - \hat{n}_y(1)) + (1 - P(1)) \hat{n}_y(1) = 0$$

$$\Rightarrow \hat{P(1)} = \frac{\hat{n}_y(1)}{n}$$ ⇒ meaning: MLE of $P(1)$

$$\hat{P}(-1) = \frac{\hat{n}_y(-1)}{n}$$

$$= \frac{\#\text{ of } y=1 \text{ data}}{\#\text{ of all data}}$$

MLE of $P(1)$ is the empirical value $\frac{\hat{n}_y(1)}{n}$

PART A $\Rightarrow$ Similarly: $\hat{P}(x_i|y) = \frac{\hat{n}_{iy}(x_i, y)}{\hat{n}_y(y)}$

have the similar interpretation

$\Rightarrow$ Conclusion:

$$\begin{cases} \hat{P}(y) = \frac{\hat{n}_y(y)}{n} \\ \hat{P}(x_i|y) = \frac{\hat{n}_{iy}(x_i, y)}{\hat{n}_y(y)} \end{cases} \Rightarrow \hat{P}(x_i, y) = \frac{\hat{n}_{iy}(x_i, y)}{n}$$

Then, $\hat{\ell}(\mathcal{D}_n) = \sum_{t=1}^{n} \log \hat{P}(x_t, y_t)$

$$= n \left[\sum_{i=1}^{d} \sum_{x_i y} \hat{P}(x_i, y) \log \hat{P}(x_i|y) + \sum_{y} \hat{P}(y) \log \hat{P}(y) \right]$$

$$= n \left[\sum_{i=1}^{d} -\hat{H}(x_i|y) + (-\hat{H}(y)) \right]$$

(Conditional) Entropy — Entropy

$$\begin{cases} H(z) = -\sum_{z} P_z(z) \log P_z(z) \\ H(x|y) = -\sum P(x, y) \log P(x|y) \end{cases}$$

conditional independent

Rmk: The maximized log-likelihood under NB model can be expressed with Conditional Entropy & Entropy.

Note: Under the null model: → feature independent with label

$$\hat{\ell}_0(\mathcal{D}_n) = n\left[\sum_{i=1}^{d} -\hat{H}(x_i) - \hat{H}(y)\right]$$

⇓

Exercise!

y

x_1 ------ x_d

Rmk: It is always true that $\hat{\ell}_{NB}(\mathcal{D}_n) \geq \hat{\ell}_0(\mathcal{D}_n)$ → model is actually better!

have more parameters

↑

Guarantee by $H(X|Y) \leq H(Y)$

Difference in $\hat{\ell}_{NB}(\mathcal{D}_n)$ & $\hat{\ell}_0(\mathcal{D}_n)$ is

$$\hat{\ell}_{NB}(\mathcal{D}_n) - \hat{\ell}_0(\mathcal{D}_n) = n\left[\sum_{i=1}^{d} \hat{H}(x_i) - \hat{H}(x_i|y)\right]$$

$$= n\sum_{i=1}^{d} \hat{I}(x_i; y) \geq 0$$

互信息 (mutual information)

- x_i discrete ✓
- x_i continuous ✗

Suggests for us to select features according to decreasing mutual information $\hat{I}(x_i; y)$ ⇒ Criterion

→ choose the bigger one

order of ⇓ Punchline!

For example, choose the biggest 4 features x_i s.t $\hat{I}(x_i; y)$ is bigger

Rmk: We can select feature x_i individually (?)

⇓

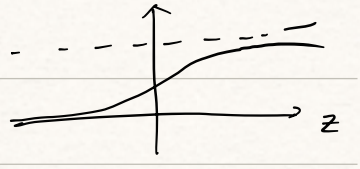
since we assumed the NB model in which x_i's are conditionally independent Given y.

→ 'BAD' Assumption!

LASSO ⇒ ℓ_1-Regularization to achieve Feature selec.

Logistic Reg. Model $P(y|x,\theta) = g(y\theta^T x)$

$g(z) = \frac{1}{1+\exp(-z)}$



(PAST)

Penalized / Regularized LR with ℓ_2 - regularization (Ridge)

⇓ $\max_{\theta} J(\theta)$ $\quad J(\theta) = \sum_{t=1}^{n} \log P(y_t|x_t,\theta) - \lambda\|\theta\|_2^2$

Claim: i) As $\lambda \to +\infty$, all $\theta_i \to 0$

ii) when $\lambda = 0$, none of $\theta_i = 0$

iii) when $\lambda \in (0, +\infty)$, none of $\theta_i = 0$

↙ Not what we want.

(NOT SPARSE) ⇓ disadvantage

if $\theta_i = 0 \Rightarrow y$ is not influenced by θ_i

⇓ Feature selection!

⇙ But Ridge cannot give us this

↓ Improvement

Method: ℓ_1 − regularization

Not smooth! ⇑

$$J(\theta) = \sum_{t=1}^{n} \log P(y_t | x_t, \theta) - \lambda \|\theta\|_1, \quad \|\theta\|_1 = \sum_{i=1}^{d} |\theta_i|$$

Claim: i) As $\lambda \to +\infty$, all $\theta_i \to 0$

ii) As $\lambda = 0$, none of $\theta_i = 0$.

iii) As $\lambda \in (0, +\infty)$, some of $\theta_i = 0$, which never happens in Ridge (ℓ_2-regularization) LR.

↓ Sparse

Two Proofs { 1. mathematical; 2. Graph.

① Mathematical Justification of Sparse

To "prove", consider the quad. loss $\ell(y, y') = \frac{1}{2}(y - y')^2$

$$\min_{\theta} \frac{1}{2n} \sum (y_t - \theta^T x_t)^2 + \lambda \|\theta\|_1$$

Square loss

Special case:

When: $n=1$, $d=1$, The object reduced to this:

$(X_1=1, y_1=y)$

$$\Downarrow$$

$$\min_{\theta} \frac{1}{2}(y_1 - \theta x_1)^2 + \lambda|\theta|$$

$$\Leftrightarrow \min_{\theta} \frac{1}{2}(y-\theta)^2 + \lambda|\theta|$$

① subgradient
② some simple method

$$\Leftrightarrow \min_{\theta} \frac{1}{2}\theta^2 - \theta y + \lambda|\theta| \quad \text{(omit const)}$$

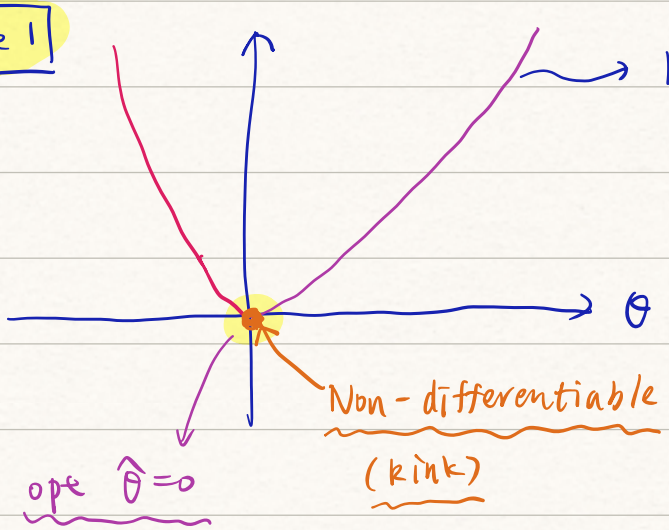
Consider 2 cases:

① opt attained at $\hat{\theta}=0$

② opt attained at $\hat{\theta}>0$

③ Actually $\hat{\theta}$ can be less than 0

Case 1

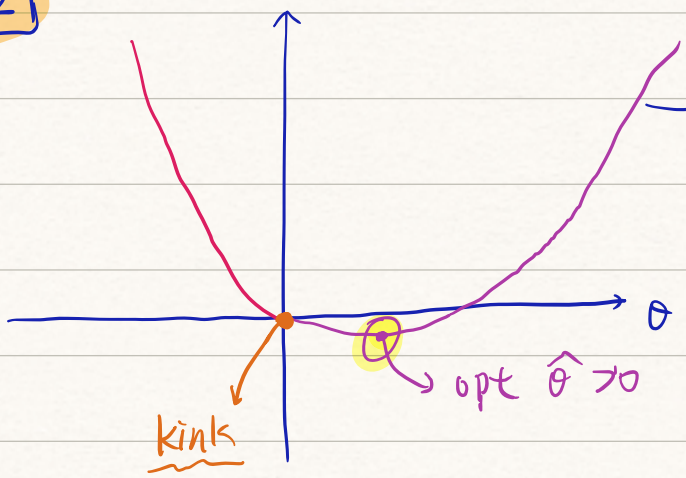


$$h(\theta) = \frac{1}{2}\theta^2 - \theta y + \lambda|\theta|$$

$$\downarrow$$

gradient $\theta - y \pm \lambda$

Case 2



$$h(\theta) = \frac{1}{2}\theta^2 - \theta y + \lambda|\theta|$$

Consider Case 1 & 2 derivative of $h(\theta)$ $\begin{cases} h'_+(0) \\ h'_-(0) \end{cases}$

$$\begin{cases} h'_+(0) = -y+\lambda \\ h'_-(0) = -y-\lambda \end{cases}$$

$y \geq \lambda$

$-y \leq -\lambda$ $\quad$ $y \geq -\lambda$

$-y \leq \lambda$

From the Figure

i) Case 1 $\Leftrightarrow$ $\begin{cases} h'_-(0) \leq 0 \\ h'_+(0) \geq 0 \end{cases}$ $\Rightarrow$ $|y| \leq \lambda$ → λ is BIG ENOUGH (relative to data)

ii) Case 2 $\Leftrightarrow$ $\begin{cases} h'_-(0) \leq 0 \\ h'_+(0) \leq 0 \end{cases}$ $\Rightarrow$ $y \geq \lambda$ → λ is relatively Small

$\Downarrow$

$h(\theta) = \frac{1}{2}\theta^2 - y\theta + \lambda\theta$ when $\theta > 0$

$\Downarrow$

$\hat{\theta} = y - \lambda$

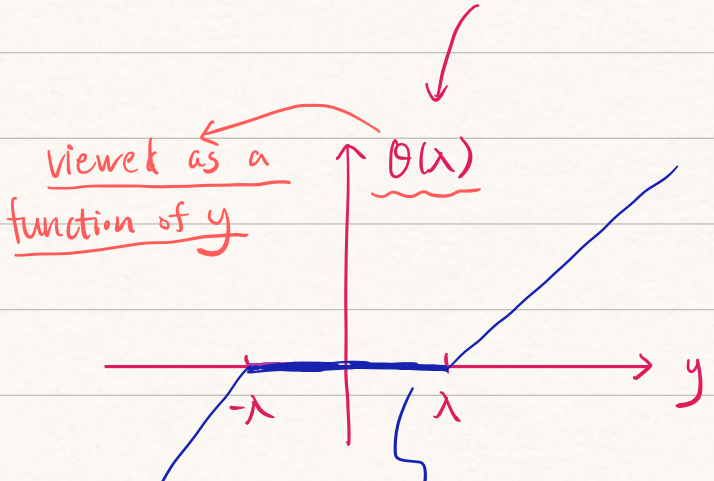
iii) Case 3 $\Leftrightarrow$ $y \leq -\lambda$ and $\hat{\theta} = y+\lambda$

☆

In Summary: our opt solution $\hat{\theta}(\lambda) = \text{sign}(y)(|y|-\lambda)_+$

$\Downarrow$

Check!



$\hat{\theta}$ sparse ZONE

Rmk: $\forall \lambda \geq |y|$ (λ BIG ENOUGH), optimal $\hat{\theta} = 0$ (sparse soln for large enough λ)

↳ soft-thresholding

② Graph (Omit) → In Notes