Lec 5

Thursday, September 12, 2019 10:52

Recap: Logistic regression (binary)

Posit $\text{logit}(\mathbb{P}(Y=1 \mid X)) = \beta^T x$

alt: $\mathbb{P}(Y=1 \mid X) = \sigma(\beta^T x)$

<u>Fitting logistic regression</u>

Logistic regrn model specifies a <u>generative</u> model for the data.

- First draw X_i
- Then compute $\sigma(\beta^T X_i)$
- $Y_i \sim \text{Bernoulli}(\sigma(\beta^T X_i))$

(i.e. generate $U_i \sim \text{Unif}[0,1]$

set $Y_i = \mathbb{I}[U_i \leq \sigma(\beta^T X_i)]$)

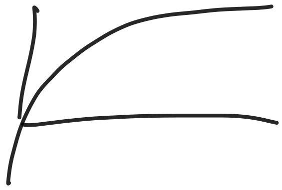
Assuming that the data is independent, for every β, there is a particular likelihood for observing our data

$$\begin{aligned}
\text{Lik}(\beta) &= \mathbb{P}(X_1, Y_1, \dots, X_n, Y_n ; \beta) \\
&= \prod_{i=1}^n \mathbb{P}(X_i, Y_i ; \beta) \\
&= \prod_{i=1}^n \mathbb{P}(Y_i \mid X_i ; \beta)\, \mathbb{P}(X_i) \\
&= \prod_{i=1}^n \mathbb{P}(X_i) \cdot \left(\begin{cases} \sigma(\beta^T X_i) & \text{if } Y_i = 1 \\ 1 - \sigma(\beta^T X_i) & \text{if } Y_i = 0 \end{cases} \right)
\end{aligned}$$

Max likelihood principle:

choose parameters that max the likelihood of observing the data

The likelihood of ...
that we observe

log:  — always increasing

$$\Rightarrow \arg\max_\beta \mathrm{Lik}(\beta)$$
$$= \arg\max_\beta \log(\mathrm{Lik}(\beta))$$
$$= \arg\min_\beta (-\log(\mathrm{Lik}(\beta))) \leftarrow \text{neg log lik}$$

$$-\log \mathrm{Lik}(\beta) = \sum_{i=1}^n \left(-\log \mathbb{P}(x_i) - \log \mathbb{P}(y_i \mid x_i ; \beta)\right)$$
$$= -\sum_i \log \mathbb{P}(x_i)$$
$$+ \sum_i \begin{cases} -\log \sigma(\beta^T x_i) & y_i = 1 \\ -\log(1-\sigma(\beta^T x_i)) & y_i = 0 \end{cases}$$

$$= -\sum_i \log(\mathbb{P}(x_i))$$
$$\underbrace{+ \sum_i \left(y_i(-\log \sigma(\beta^T x_i)) + (1-y_i)(-\log(1-\sigma(\beta^T x_i)))\right)}_{\mathcal{L}(\beta) \leftarrow \text{neg log lik fun}}$$

$$\arg\max_\beta \mathrm{Lik}(\beta) = \arg\min_\beta \mathcal{L}(\beta)$$

$$-\log(\sigma(\beta^T x_i)) = -\log\left(\frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}\right) = -\beta^T x_i + \log(1+e^{\beta^T x_i})$$

$$-\log(1-\sigma(\beta^T x_i)) = -\log\left(\frac{1}{1+e^{\beta^T x_i}}\right) = +\log(1+\beta^T x_i)$$

$$\mathcal{L}(\beta) = \sum_i \Big(\cancel{y_i} \log(1+e^{\beta^T x_i}) - y_i \beta^T x_i$$
$$+ (1-\cancel{y_i}) \log(1+e^{\beta^T x_i})\Big)$$

$$= \sum_i \left(\log(1+e^{\beta^T x_i}) - y_i \beta^T x_i \right)$$

Fitting logistic regrn: solving $\min_\beta \ell(\beta)$

call $\hat{\beta}$ the optimal β

Can solve w/ $\nabla \ell(\beta) = 0$

but computationally more difficult than OLS

Revisit after fall break

More than two categories: Multinomial Logistic Regression

$Y \in \{1, \dots, m\}$ $\quad m > 2$

Posit $\mathbb{P}(Y=j \mid X=x) \propto e^{\beta_j^T x}$

$\beta_1, \dots, \beta_m \in \mathbb{R}^m$

$$1 = \sum_j \mathbb{P}(Y=j \mid X=x) \Rightarrow \mathbb{P}(Y=j \mid X=x) = \frac{e^{\beta_j^T x}}{\sum_{j'=1}^m e^{\beta_{j'}^T x}}$$

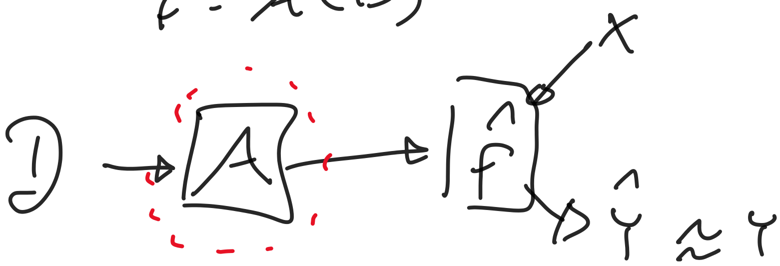
Fix $\beta_m = \vec{0}$.

Fit by maximum likelihood

The bias-variance tradeoff

A supervised learning algo can be understood as a fn A from $D = \{x_1, y_1, \dots, x_n, y_n\}$ to a prediction rule $\hat{f}$

$$\hat{f} = \mathcal{A}(D)$$



What makes a good algo?

A good algo should do well over all over all dataset D

$$R(\mathcal{A}) = \mathbb{E}_D R(\mathcal{A}(D))$$

$$= \mathbb{E}_D[R(\hat{f})]$$

$$= \mathbb{E}_D\Big[\mathbb{E}_{X,Y}[\ell(Y, \hat{f}(X))]\Big]$$

random b/c new random test pt

random b/c new random training datasets

for regression

$$= \mathbb{E}_D\Big[\underbrace{\mathbb{E}_{X,Y}\big[(Y - \hat{f}(X))^2\big]}_{\circledast}\Big]$$

$$\circledast = \mathbb{E}_{X,Y}[(Y - f(X))^2]$$

$$= \mathbb{E}_X\big[\mathbb{E}_Y[(Y - f(X))^2 \mid X]\big]$$

$$= \mathbb{E}_X\big[\mathbb{E}_Y[Y^2 - 2Yf(X) + f(X)^2 \mid X]\big]$$

$$= \mathbb{E}_X\big[\mathbb{E}_Y[Y^2 \mid X] - 2f(X)\underbrace{\mathbb{E}[Y \mid X]}_{f^*(X)} + f(X)^2\big]$$

$$= \mathbb{E}_X\Big[\underbrace{\mathbb{E}_Y[Y^2|X] - f^*(x)^2}_{Var(Y|X)} + \underbrace{f^*(x)^2 - 2f(x)f^*(x) + f(x)^2}_{(f^*(x)-f(x))^2}\Big]$$

$$= \underbrace{\mathbb{E}_X\left[(f^*(x)-f(x))^2\right]}_{\substack{\text{reducible error}\\ \text{can be made}\\ \text{small w/ } f \approx f^*}} + \underbrace{\mathbb{E}_X \, Var(Y|X)}_{\substack{\text{irreducible error}\\ \text{no matter what } f \text{ is}}}$$

Hence $R(\mathcal{A}) = \mathbb{E}_D[\quad]$

$$= \mathbb{E}_X \, Var(Y|X)$$

$$+ \, \mathbb{E}_X\Big[\underbrace{\mathbb{E}_D\left[(f^*(x) - \hat{f}(x))^2\right]}_{Err(x)}\Big]$$

$$Err(x) = \mathbb{E}_D\left[(\hat{f}(x) - f^*(x))^2\right]$$

$$= \mathbb{E}_D\left[\left((\hat{f}(x) - \mathbb{E}_D\hat{f}(x)) + (\mathbb{E}_D\hat{f}(x) - f^*(x))\right)^2\right]$$

$$= \mathbb{E}_D\left[(\hat{f}(x) - \mathbb{E}_D\hat{f}(x))^2\right] + (\mathbb{E}_D\hat{f}(x) - f^*(x))^2$$

$$+ 2\left(\cancel{\mathbb{E}_D[\hat{f}(x) - \mathbb{E}_D\hat{f}(x)]}^{\,0}\right)(\mathbb{E}_D\hat{f}(x) - f^*(x))$$

$$= Var_D \hat{f}(x) + \underbrace{\Big(\underbrace{\mathbb{E}_D\hat{f}(x) - f^*(x)}_{\text{bias}}\Big)^2}_{\text{bias}^2}$$