Lecture 7 - linear & logistic regression

Thursday, September 26, 2024 2:37 PM

today: linear regression
logistic "

(asymptotic) properties MLE:

under suitable regularity conditions on Θ & $p(x;\theta)$ $\quad \hat{\theta}_n = \arg\max_\theta \frac{1}{n}\sum_{i=1}^n \log p(x_i;\theta)$

a) $\hat{\theta}_n \xrightarrow{p} \theta$ "consistent"

b) CLT (central limit theorem) $\quad \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{dist.} N(0, I(\theta)^{-1/2})$

$[D_n \sim p_\theta^{\otimes n}]$

$I(\theta)$: information matrix

c) asymptotically optimal (Cramer-Rao lower bound)
i.e. it has minimal asymptotic scaled variance among all "reasonable" estimators

(always true) (not just asymptotic) b.d) **invariance**: MLE is preserved under reparametrization

suppose have a bijection $f: \Theta \to \Theta'$

then $\widehat{f(\theta)} = f(\hat{\theta})$

example: $\widehat{(\sigma^2)} = (\hat{\sigma})^2$

since $\sigma^2 = \sigma \cdot \sigma$, $\hat{\sigma}^2 = \hat{\sigma} \cdot \hat{\sigma}$

* if not a bijection, can generalize the MLE with "profile likelihood"

suppose $g: \Theta \to \Lambda$

profile likelihood $\triangleq L(\eta) = \max_{\theta: g(\theta)=\eta} p(\text{data};\theta)$

define $\hat{\eta}_{MLE} \triangleq \arg\max_{\eta \in \Lambda} L(\eta)$

then we have $\boxed{\hat{\eta}_{MLE} = g(\hat{\theta}_{MLE})}$ "plug in estimate"

$N(\mu, \sigma^2)$

e.g. $g(\mu) = \mu^2$

prediction

want to learn a prediction fct, $h: \mathcal{X} \to \mathcal{Y}$

$x \in \mathbb{R}^d$

$\mathcal{Y} = \{0,1\}$ → binary classification
$\mathcal{Y} = \{0,1,\dots,K-1\}$ → multiclass classification

$\mathcal{Y} = \mathbb{R}$ → regression

$p(x,y) = \underbrace{p(y|x)}_{\text{"prediction model"}} \underbrace{p(x)}_{\text{marginal model on } X}$

"prediction model" — regressor

$$p(x,y) = \underbrace{p(y|x)}\ \underbrace{p(x)}$$

marginal model on X

$$= \underbrace{p(x|y)}\ p(y)$$

"class conditional" — prior over class

"generative perspective" (in context of classification) → model $p(x)$ as well

vs.

"conditional perspective" → only model $p(y|x)$

"more discriminative" ↔ traditionally called "discriminative"

generative	conditional	"fully discriminative"
model $p_\theta(x,y)$	model $p_\theta(y\|x)$	model $h_\theta : \mathcal{X} \to \mathcal{Y}$
MLE	max conditional likelihood (MCL)	(not nec. derived from $p(y\|x)$) reg. ERM; etc. $\frac{1}{n}\sum_{i=1}^{n} \ell(y^{(i)}, h_\theta(x^{(i)})) + \text{reg}(\theta)$
more assumptions ⇒ less robust predictions		less assumption more robust

$$\hat{h}(x) \triangleq \operatorname*{argmin}_{\tilde{y}\in\mathcal{Y}} \sum_{y} \hat{p}_\theta(y|x)\, \ell(y,\tilde{y})$$

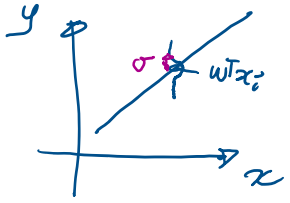
if $\ell(y,\tilde{y}) = \mathbb{1}\{y \neq \tilde{y}\}$ (0-1 loss) then $\hat{h}(x) = \operatorname*{argmax}_{\tilde{y}\in\mathcal{Y}} p(\tilde{y}|x)$

linear regression: derive/motivate with conditional approach to regression ($Y \in \mathbb{R}$)

$$p(y|x; w) = \mathcal{N}(y | \langle w, x\rangle, \sigma^2)$$

w: parameter; $\langle w, x\rangle = w^T x$

$x \in \mathbb{R}^d$, $w \in \mathbb{R}^d$



equivalently: $Y_i = w^T X_i + \varepsilon_i$

where $\varepsilon_i | X_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

[aside: we use "offset" notation for x

i.e. $x = \begin{pmatrix} \tilde{x} \\ 1 \end{pmatrix}$ $\tilde{x} \in \mathbb{R}^{d-1}$ ← "constant feature"

thus $\langle w, x\rangle = \langle w_{1:d-1}, \tilde{x}\rangle + w_d$ ("bias" / "offset") (usually denoted as b)]

* dataset $(x_i, y_i)_{i=1}^n$ $X_i \sim$ whatever (don't care)

$Y_i | X_i \overset{\text{indep}}{\sim} \mathcal{N}(w^T x_i, \sigma^2)$

conditional likelihood:

$Y_i | X_i \sim N(w^T x_i, \sigma^2)$

conditional likelihood:

$$p(y_{1:n} \mid x_{1:n}) \overset{\text{indep}}{=} \prod_{i=1}^n p(y_i | x_i)$$

this is not a concave fct. of σ^2 ($f(u) = -\frac{a}{u} - b\log u$)

$$\log[\quad] = \sum_{i=1}^n \left[-\frac{(y_i - w^T x_i)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \right]$$

$\sigma^2 = \frac{1}{\lambda}$ $\quad \lambda = \frac{1}{\sigma^2}$ $\quad f(\lambda) = \underbrace{-a\lambda}_{\text{concave}} + \underbrace{b\log\lambda}_{\text{concave}}$

$$\frac{\partial}{\partial \lambda}(\quad) = 0 \Rightarrow \sum_{i=1}^n \left(-\frac{1}{2}(y_i - w^T x_i)^2 + \frac{1}{2}\frac{1}{\lambda} \right) = 0$$

$$\Rightarrow \boxed{\hat{\lambda}_{MLE} = \left[\frac{1}{n} \sum_{i=1}^n (y_i - w^T x_i)^2 \right]^{-1}}$$ (global max by concavity)

(by invariance of MLE) $\Rightarrow \boxed{\hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^n (y_i - w^T x_i)^2}$

$\Rightarrow$ conclude that this correct global in σ^2 for w fixed

Ishy "design matrix" (matrix???)

$$X \triangleq \begin{pmatrix} - x_1^T - \\ - x_2^T - \\ \vdots \\ - x_n^T - \end{pmatrix} \quad n \times d \text{ matrix} \qquad y \triangleq \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$Xw = \begin{pmatrix} x_1^T w \\ x_2^T w \\ \vdots \\ x_n^T w \end{pmatrix} \in \mathbb{R}^{n \times 1} \qquad \sum_{i=1}^n (y_i - w^T x_i)^2 = \|y - Xw\|_2^2$$

can rewrite $-\log p(y_{1:n} \mid X) = \frac{\|y - Xw\|^2}{2\sigma^2} + \text{fct.}(\sigma^2)$ (X: design matrix)

MCL $\longrightarrow \min_w \|y - Xw\|_2^2 \iff$ projection of y on the column space of design matrix X

$$Xw = \sum_{j=1}^d X_{:j} w_j$$ (j^{th} column of X)

$\hat{w}_{MLE} = \underset{w \in \mathbb{R}^d}{\text{argmin}} \|y - Xw\|_2^2$ "least square"

algebra: want $\nabla_w \overset{\text{set to}}{=} 0$

$$\nabla_w \left[(y - Xw)^T (y - Xw) \right] \overset{\text{want}}{=} 0$$

$$\nabla_w \left[\|y\|^2 - 2y^T Xw + w^T X^T X w \right]$$

vector $\nabla_w (w^T A w) = (A + A^T) w$

$\nabla_w [\dots] = (A + A^T) w$

$\nabla_w [\|y\|^2 - 2y^T X w + w^T X^T X w]$

$0 - 2X^T y + 2X^T X w = 0$

$\Rightarrow \boxed{(X^T X) w^* = X^T y}$ "normal equation"

$\|\cdot\|^2$ convex fct. of w $\Rightarrow$ stat. pt. are global min

a) if $X^T X$ ($d \times d$) is invertible, then have unique soln

$\hat{w}_{MLE} = \underbrace{(X^T X)^{-1} X^T}_{X^+} y$

X is $n \times d$ $\Rightarrow \text{rank}(X) \le \min\{n, d\}$

$\text{rank}(X^T X) = \text{rank}(X)$

$X^T X$ is invertible $\Rightarrow \boxed{n \ge d}$ $\le \min\{n, d\}$

prediction on training set: $\hat{y} = X\hat{w} = \underbrace{X (X^T X)^{-1} X^T}_{\text{projection on column space of } X} y$

↝ recall geometric perspective?

(*) if $n < d$ (i.e. high dimension or low data regime) then $X^T X$ is not invertible

b) if $(X^T X)$ is not invertible → there is no unique soln

any $\hat{w}$ s.t. $(X^T X)\hat{w} = X^T y$ is a MCL soln

could choose $\hat{w} = \underset{w}{\text{argmin}} \|w\|_2 = X^+ y$ ← Moore-Penrose pseudo-inverse

s.t. $(X^T X) w = X^T y$

$X^+ = (X^T X)^{-1} X^T$ when X is full rank

SVD

$X = U \Sigma V^T$ $\quad U^T U = I_n \quad V V^T = I_d$

X: $n \times d$, U: $n \times n$, Σ: $n \times d$, V^T: $d \times d$

$\Sigma = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ & & \sigma_d \\ 0 & & \end{pmatrix}$ $\quad X^+ = V \Sigma^+ U^T$

$\Sigma^+ = \begin{pmatrix} \sigma_1^+ & & 0 \\ & \ddots & \\ & & \sigma_d^+ \\ 0 & & \end{pmatrix}$

$\sigma_i^+ = \begin{cases} 1/\sigma_i & \text{if } \sigma_i \neq 0 \\ 0 & \text{o.w.} \end{cases}$

problem: pseudo-inverse is not numerically stable

instead it is better to regularize to get similar effect

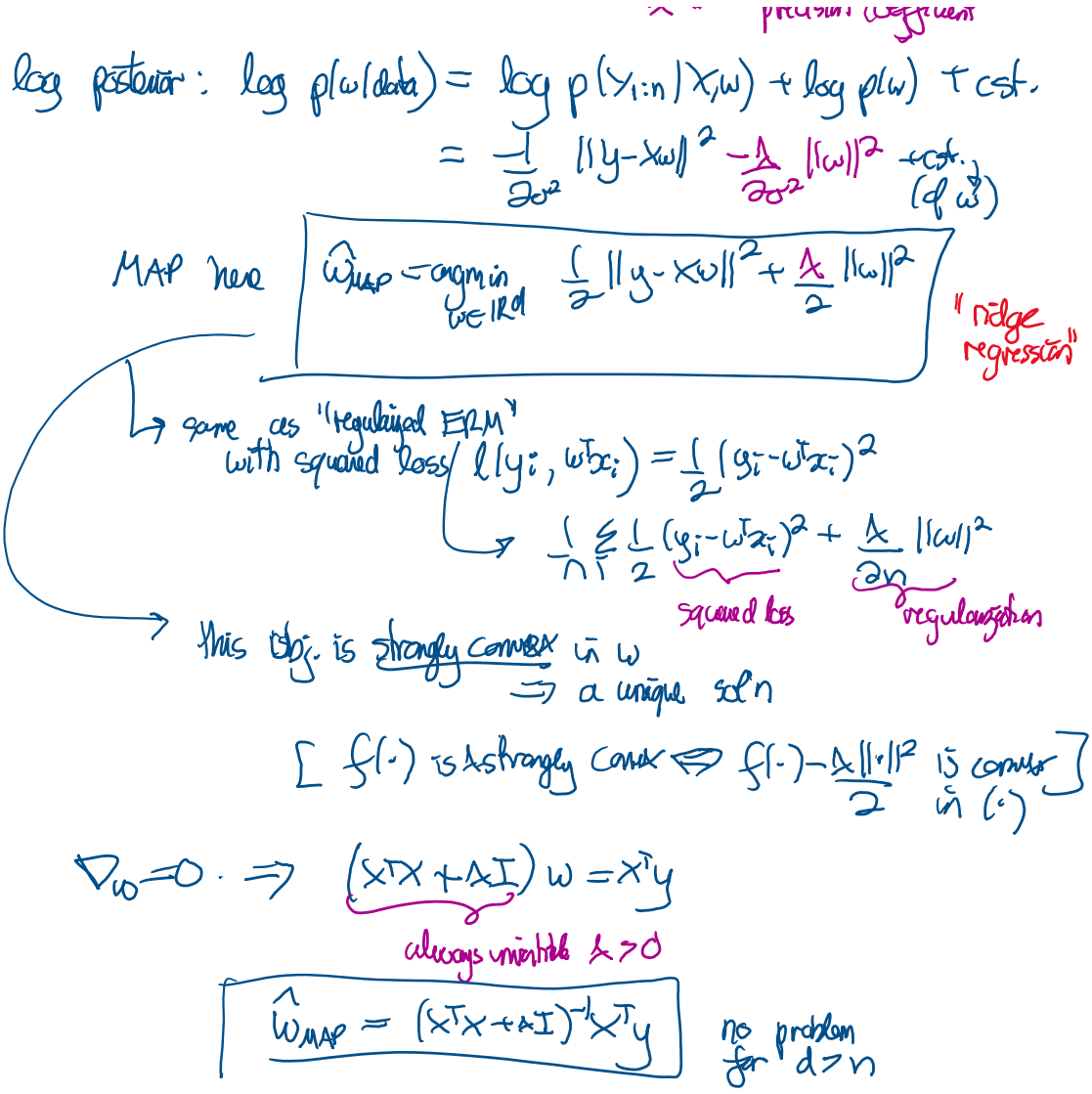
→ $\hat{w}_{MAP}(\lambda) \xrightarrow{\lambda \to 0} \hat{w}_{\text{pseudo-inverse}}$

regularization: (can be motivated from MAP pt. of view)

suppose we put a prior $p(w) = N(w | 0, \sigma^2 I)$ ($d \times d$ identity matrix)

$\lambda \leftarrow$ "precision coefficient"

log posterior: $\log p(w | \text{data}) = \log p(y_{1:n} | X, w) + \log p(w) + \text{cst.}$



- **note about sigma^2 being a global max**

  (**aside**: showing that the sigma^2 above is the **global max** is subtle because the objective is not concave in sigma^2. I give more info here for your curiosity, but it is not required for the assignment. And it's easier instead to just use lambda= 1/sigma^2 instead as the parameterization as I have done, as the objective is then concave in lambda...)

  - Formally, to find a global max of a *differentiable objective*, you need to check all **stationary points** (zero gradient points), **as well as the values at the boundary of the domain.**

    Thus here, you would need to show that the objective cannot take higher value anywhere at the boundary of the domain (which is the case here (exercise!), as the objective goes to -infinity at the boundary), so you are done (this is the only possible global optimum -- a maximum here, as it should be, given that there are no other stationary points and all values are lower at the boundary, but one could also explicitly check the Hessian to see that it is strictly negative definite at the stationary point, i.e. it looks like a local maximum).

    Note that we will see later in the class that the Gaussian is in the exponential family, with a log-concave likelihood in the right ("natural") parameterization, and thus using the invariance principle of the MLE, we could also easily deduce the MLE in the "moment" parameterization which is the usual (mu,sigma^2) one, without having to worry about local optima...

  - for a cute counter-example illustrating that a differentiable function could have only one stationary point which is a local min but *not a global min* (and thus why one need to look at the values at the boundary), see:
    - https://en.wikipedia.org/wiki/Maxima_and_minima#Functions_of_more_than_one_variable
    - i.e.

$$f(x,y) = x^2 + y^2(1-x)^3, \qquad x, y \in \mathbb{R},$$

shows. Its only critical point is at (0,0), which is a local minimum with $f(0,0) = 0$. However, it cannot be a global one, because $f(2,3) = -5$.

(see picture of function here)

(and note that the "Mountain pass theorem" which basically says that if you have a strict local optimum with another point somewhere with the same value, then there must be a saddle point somewhere (a "mountain pass") i.e. another stationary point, **does not hold for this counter-example** as one of the required regularity condition, the "Palais-Smale compactness condition" fails. Here, the saddle point (which should intuitively exist) "happens at infinity", which is why it only has one stationary point despite (0,0) not being a global minimum)

- the moral of the story: intuitions for multivariate optimization are often misleading! (this counter-example would not work in 1d because of Rolle's theorem)