Lecture 7 - scribbles

Tuesday, September 22, 2020 14:29

today: • finish estimators
• linear regression

other example of estimator

4) in the context of prediction $\mathcal{A} = \{ (f: \mathcal{X} \to \mathcal{Y}) \}$ $\mathcal{X}$ = input space, $\mathcal{Y}$ = output

example of $\delta: \mathcal{D} \to \mathcal{A}$

is using empirical "risk" minimization (ERM)

↳ empirical risk i.e. generalization error

i.e. $L(P, f) = \mathbb{E}_{(X,Y) \sim P} [\ell(Y, f(X))]$

replace with $\hat{\mathbb{E}} [\ell(Y, f(X))] = \frac{1}{n} \sum_{i=1}^{n} \ell(y^{(i)}, f(x^{(i)}))$

$\hat{f}_{ERM} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \hat{\mathbb{E}} [\ell(Y, f(X))]$

↑ hypothesis class

James-Stein estimator:

estimator to estimate the mean of $N(\vec{\mu}, \sigma^2 I)$ ← d independent Gaussian variables $X_i \overset{\text{indep}}{\sim} N(\mu_i, \sigma^2)$

δ_{JS} is biased, but much lower variance than MLE

recall bias-variance decomposition for squared loss $R(\theta, \hat{\theta}) = \mathbb{E} \|\theta - \hat{\theta}\|_2^2$

$= \underbrace{\|\mathbb{E}\hat{\theta} - \theta\|_2^2}_{\text{bias}} + \underbrace{\mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]}_{\text{variance}}$

δ_{JS} actually strictly dominates δ_{MLE} for $d \geq 3$ ← dimension of $\vec{\mu}$

i.e. $R(\theta, \delta_{JS}) \leq R(\theta, \delta_{MLE})$ $\forall \theta$
and $\exists \theta$ s.t. $R(\theta, \delta_{JS}) < R(\theta, \delta_{MLE})$

→ MLE is inadmissible in this case

(can interpret the δ_{JS} as an "empirical" Bayesian method)

(asymptotic) properties of MLE:

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

a) $\hat{\theta}_n \xrightarrow{P} \theta$ "consistent" $\quad$ [$D_n \sim p_\theta^{\otimes n}$

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

$\mathbb{E}\|\hat{\theta}_n - \theta\|^2 \to 0$, $\hat{\theta}_n \xrightarrow{L^2} \theta$, $\hat{\theta}_n \xrightarrow{P} \theta$

c) asymptotically optimal (Cramer-Rao lower bound) (scaled)

i.e. it has minimal asymptotic variance among all "reasonable" estimators

d) invariance : MLE is preserved under reparameterization

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

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

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

$\widehat{\sin}\,\sigma^2 = \sin\hat{\sigma}^2$

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

Suppose $g: \Theta \to \Lambda$

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

define $\hat{\eta}_{MLE} \triangleq \underset{\eta\in\Lambda}{\operatorname{argmax}}\, L(\eta)$

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

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

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

15h34

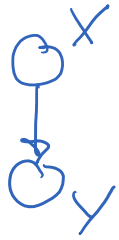
prediction

want to learn prediction $f: \mathcal{X} \to \mathcal{Y}$ $\quad x \sim S_{n+1} \to$ learn f

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

$\mathcal{Y} = \{0,1\}$ → binary class.
$\mathcal{Y} = \{0,1,\dots,K-1\}$ → multiclass
$\mathcal{Y} = \mathbb{R}$ → regression

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



"prediction model" — model over $\mathcal{X}$

$$p(x,y) = p(y|x)\,p(x)$$
$$= p(x|y)\,p(y)$$

"class conditional" — prior over class

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

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

"more discriminative"

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

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

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

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

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

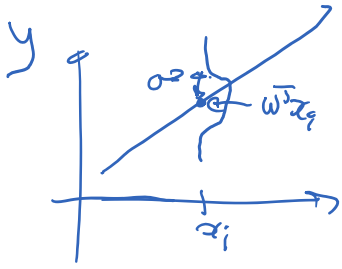
$\langle w, x \rangle = w^T x$

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

$p(y|x;w) = N(y|w^T x, \sigma^2)$ $\quad N(\mu,\sigma^2)$ $\quad N(y|\mu,\sigma^2)$

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

equivalently: $Y_i = w^T X_i + \varepsilon_i$ where $\varepsilon_i | X_i \overset{iid}{\sim} N(0,\sigma^2)$



[aside: we use "offset" notation for x

ie. $x = \begin{pmatrix} \tilde{x} \\ 1 \end{pmatrix}$ $\quad \tilde{x} \in \mathbb{R}^{d-1}$ ← "constant feature" (bias/offset)

thus $\langle w, x \rangle = \langle w_{1:d-1}, \tilde{x} \rangle + w_d$]

- dataset $(x_i, y_i)_{i=1}^n$ $\quad X_i \sim$ whatever

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

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

$\log(\quad) = \sum_{i=1}^n \left[-\frac{(y_i - w^T x_i)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \right]$ — not a concave fct of σ^2

$\frac{\partial}{\partial(\sigma^2)}(\quad) = 0 \rightarrow \sum_{i=1}^n \left[-(y_i - w^T x_i)^2 \left(\frac{-1}{(\sigma^2)^2}\right) - \frac{1}{2}\frac{1}{\sigma^2} \right] = 0$

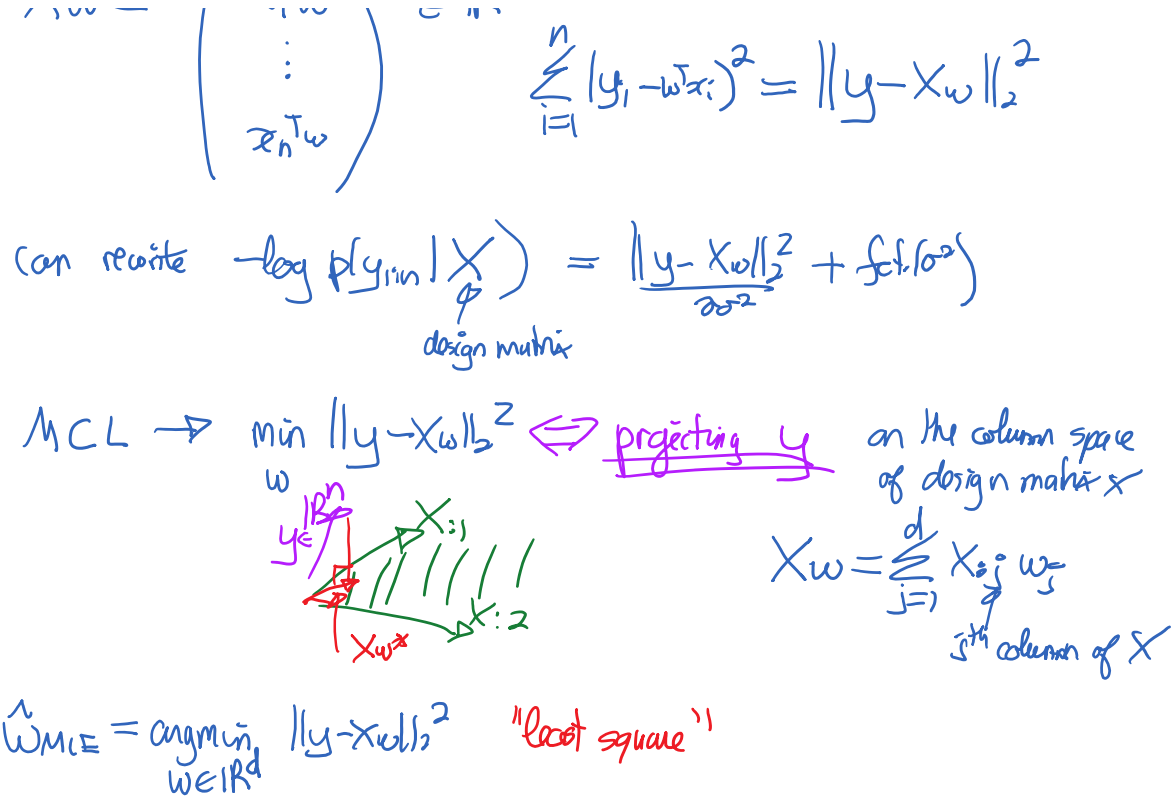
$\Rightarrow \boxed{\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^n (y_i - w^T x_i)^2}$

obj $\to -\infty$ as $\sigma \to 0$ or $\sigma \to +\infty$ — so conclude that this is correct global max for w fixed

"design matrix" $X \triangleq$ matrix (???) $n \times d$ matrix $\triangleq \begin{pmatrix} -x_1^T- \\ \vdots \\ -x_n^T- \end{pmatrix}$ $\quad$ vector $y \triangleq \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$

$Xw = \begin{pmatrix} x_1^T w \\ \vdots \end{pmatrix} \in \mathbb{R}^{n \times 1}$

$\sum_{i=1}^n (y_i - w^T x_i)^2 = \|y - Xw\|_2^2$



"Bayesian Methods for Hackers" -- cute book on practical Bayesian approach with probabilistic programming: https://camdavidsonpilon.github.io/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/

(thanks to Dora Jambor for the reference!)

- **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.)

  - 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)