Lecture 6 - scribbles

Friday, September 22, 2017
13:21

today: • properties of estimators
• gen. vs. disc. methods
• linear & logistic regression

properties of estimators: bias-variance

iid setting: $P = p_\theta^{\otimes n}$ $\quad D_n = (X_i)_{i=1}^n$

$X_i \overset{iid}{\sim} p_\theta$

notation $\hat{\theta}_n = \delta_n(D_n)$

↳ to emphasize dependence on n

study $R(\theta, \delta_n)$ as a function of n

in particular, would like $R(\theta, \delta_n) \xrightarrow{n\to\infty} 0$

"consistency"

for estimation, typical loss: squared loss $L(\theta, \delta_n(D)) = \|\theta - \delta_n(D)\|_2^2$

(standard statistical consistency: $\hat{\theta}_n \xrightarrow{P} \theta$ "in probability"

i.e. $\forall \varepsilon > 0$, $P\{\|\hat{\theta}_n - \theta\| > \varepsilon\} \xrightarrow{n\to\infty} 0$

↳ randomness is from $\hat{\theta}_n = \delta_n(D)$ ↳ random)

risk $R(\theta, \delta_n) = \mathbb{E}_{D\sim P}\left[\|\theta - \hat{\theta}_n\|_2^2\right]$

$= \mathbb{E}\left[\|\theta - \mathbb{E}[\hat{\theta}_n] + \mathbb{E}[\hat{\theta}_n] - \hat{\theta}_n\|^2\right]$

$= \mathbb{E}\left[\|\theta - \mathbb{E}[\hat{\theta}_n]\|^2\right] + \mathbb{E}\left[\|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\|^2\right]$

$$= \mathbb{E}[\|\theta - \mathbb{E}[\hat{\theta}_n]\|^2] + \mathbb{E}[\|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\|^2]$$

$$+ 2\,\mathbb{E}[\langle \theta - \mathbb{E}[\hat{\theta}_n],\ \mathbb{E}[\hat{\theta}_n] - \hat{\theta}_n \rangle]$$

const.

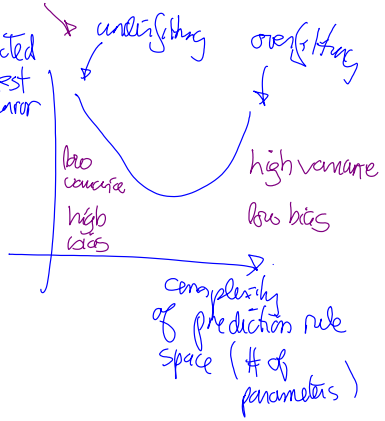
$$2\langle \theta - \mathbb{E}[\hat{\theta}_n],\ \mathbb{E}(\mathbb{E}[\hat{\theta}_n] - \hat{\theta}_n) \rangle \to 0$$

$$R(\theta, \delta_n) = \mathbb{E}_{Data}[\|\theta - \hat{\theta}_n\|_2^2] = \|\theta - \mathbb{E}[\hat{\theta}_n]\|^2 + \mathbb{E}[\|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\|^2]$$

$\triangleq$ bias variance($\hat{\theta}_n$)

> risk for squared loss = $\|\text{bias}\|^2$ + variance($\hat{\theta}_n$)

bias-variance decomposition tradeoff



James-Stein estimator: for estimating mean of $N(\vec{\mu}, \sigma^2 I)$

↳ is biased, but lower variance than MLE

and actually, J-S estimator dominates MLE for $d \geq 3$

ie. $R(\theta, \delta_{JS}) \leq R(\theta, \delta_{MLE})$

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

MLE is sometimes inadmissible?

consistency: $(\hat{\theta}_n \xrightarrow{?} \theta)$

also mentioned $R(\theta, \delta_n) \xrightarrow{n\to\infty} 0$ $\mathbb{E}[\|\hat{\theta}_n - \theta\|^2] \to 0$ "convergence in ℓ_2"

convergence in ℓ_2 $\Rightarrow$ convergence in prob.

* bias $\to 0$ and variance $\to 0$ $\Rightarrow$ consistency

note: in hwk 1, for simplicity, by consistency, I mean that R(theta,delta_n) -> 0 (so no need to worry about the subtleties of convergence of probability to show that an estimator is not consistent)

variance → 0

properties (asymptotic) of MLE

under regularity conditions on $\{p(x;\theta)\}$

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

b) CLT: $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, I(\theta)^{-1/2})$

$I(\theta)$: information matrix

c) asymptotically optimal (Cramer-Rao lower bound)

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

↳ · consistent

· ...

d) invariance: MLE is preserved under reparameterization

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

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

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

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

define $\hat{\eta}_{MLE} \triangleq \operatorname{argmax}_{\eta \in g(\Theta)} L(\eta)$

then we have $\hat{\eta} = g(\hat{\theta}_{MLE})$ "invariance"

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

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

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

$L_2 \Rightarrow P \rightarrow \Rightarrow g$

$as, \Rightarrow$

---

prediction:

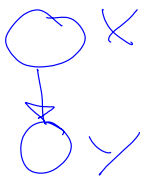
want learn a prediction function $h: \mathcal{X} \rightarrow \mathcal{Y}$

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

$\mathcal{Y} = \{0,1\} \rightarrow$ binary classification

$\{0,1,\dots,k\} \Rightarrow$ multiclass "

$\mathbb{R} \rightarrow$ regression



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

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

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

"class-conditional" prior over classes

generative perspective → model $p(x)$ as well

conditional perspective → only models $p(y|x)$

(traditionally called "discriminative"

more disc.

gen	conditional	"fully discrim"
model $p_\theta(x,y)$	model $p_\theta(y\|x)$	model $h_\theta : X \to Y$ (not nec. $p(y\|x)$) use $\ell(y, y')$ for estimation
more assumptions $\Rightarrow$ less robust for prediction		more robust

Linear regression: conditional approach to regression ($Y \in \mathbb{R}$)

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

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

parameter: w

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

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

equivalently: $Y = w^T X + \varepsilon$ where $\varepsilon \overset{\text{indep.}}{\sim} N(0, \sigma^2)$

[aside: we'll use "offset" notation for x i.e. $x = \begin{pmatrix} \tilde{x} \\ 1 \end{pmatrix}$, $\tilde{x} \in \mathbb{R}^{d-1}$; $1 \leftarrow$ "constant feature"]

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

w_d: bias/offset

- dataset $(x_i, y_i)_{i=1}^n$

$X_i \sim$ whatever

$Y_i \mid X_i \overset{\text{iid}}{\sim} N(w^T X_i, \sigma^2)$

$$N(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

conditional likelihood $p(y_1, \dots, y_n \mid x_1, \dots, x_n) = \prod_{i=1}^n p(y_i \mid x_i)$

log ... $= \log p(y_1, \dots \mid x \dots) = \sum^n [(y_i - w^T \dots)^2 \dots \log(2\pi\sigma^2)]$

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

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

(see note below about sigma^2 being true global max)

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

"design matrix"

$\underset{n \times d}{\overleftrightarrow{X}} = \begin{pmatrix} — x_1^T — \\ \vdots \\ — x_n^T — \end{pmatrix} \in \mathbb{R}^{n \times d}$ $\quad \vec{y} \triangleq \begin{pmatrix} y_1 \\ y_i \\ y_n \end{pmatrix}$

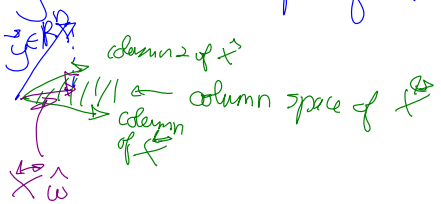
$\overleftrightarrow{X} w = \begin{pmatrix} \vdots \\ x_i^T w \\ \vdots \end{pmatrix} \in \mathbb{R}^{n \times 1}$ $\quad \| \vec{y} - \overleftrightarrow{X} w \|_2^2 = \sum_i (y_i - w^T x_i)^2$

can rewrite:

$-\log p(\vec{y} | \overleftrightarrow{X}) = \frac{\| \vec{y} - \overleftrightarrow{X} w \|_2^2}{2\sigma^2} + \text{function}(\sigma^2)$

minimizing $\| \vec{y} - \overleftrightarrow{X} w \|^2 \rightsquigarrow$ projecting $\vec{y}$ on the column space of $\overleftrightarrow{X}$

(geometric view)



$\overleftrightarrow{X} w = \sum_{j=1}^{d} \overleftrightarrow{X}_{:j} w_j$

j^{th} column of $\overleftrightarrow{X}$

- **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)