Cross-validation and Splines

Vahid Partovi Nia

Advanced Machine Learning: Lecture 04

March 3, 2018

Outline

Likelihood

- Why do we need parametric models?
- Why do we use likelihood?
- Why maximum likelihood is good?
- What information means?
- How information is related to data?

$\mathbb{KL}$ divergence between the assumed class $f(x \mid \theta)$ from true data distribution $f(x \mid \theta_0)$ is

$$\begin{aligned}\mathbb{KL}(\theta_0, \theta) &= \int \log\left\{\frac{f(x \mid \theta_0)}{f(x \mid \theta)}\right\} f(x \mid \theta_0) \\ &= \mathbb{E}_{\theta_0}\left\{\frac{f(x \mid \theta_0)}{f(x \mid \theta)}\right\}\end{aligned}$$

$\mathbb{KL}$ divergence between the assumed class $f(x \mid \theta)$ from true data distribution $f(x \mid \theta_0)$ is

$$\begin{aligned} \mathbb{KL}(\theta_0, \theta) &= \int \log\left\{\frac{f(x \mid \theta_0)}{f(x \mid \theta)}\right\} f(x \mid \theta_0) \\ &= \mathbb{E}_{\theta_0}\left\{\frac{f(x \mid \theta_0)}{f(x \mid \theta)}\right\} \end{aligned}$$

$$\mathbb{KL}(\theta_0, \theta) \neq \mathbb{KL}(\theta, \theta_0)$$

Cross entropy of the assumed class $f(x \mid \theta)$ from true data distribution $f(x \mid \theta_0)$ is

$$\mathbb{H}(\theta, \theta_0) = \int \log f(x \mid \theta) f(x \mid \theta_0) dx$$

KL and H

$$\mathbb{KL}(\theta_0, \theta) = \mathbb{H}(\theta_0, \theta_0) - \mathbb{H}(\theta, \theta_0)$$

- $\mathbb{KL}(\theta_0, \theta) > 0$ iff $f(x \mid \theta_0) \neq f(x \mid \theta)$ on a set of x with positive measure.
- $\mathbb{KL}(\theta_0, \theta) = 0$ iff $f(x \mid \theta_0) = f(x \mid \theta)$ almost everywhere.
- $\mathbb{KL}_n(\theta_0, \theta) = n\mathbb{KL}(\theta_0, \theta)$ for a set of i.i.d observations $(x_1, \ldots, x_n)$.
- $\frac{\partial \mathbb{H}(\theta,\theta_0)}{\partial \theta}\big|_{\theta=\theta_0} = 0$
- $\frac{\partial^2 \mathbb{H}(\theta,\theta_0)}{\partial \theta \partial \theta^\top}\big|_{\theta=\theta_0} = -J(\theta_0)$ where $J(.)$ is the observed information.

More about entropy

Suppose $A = \{A_1, \ldots, A_k\}$ with probabilities $p_1, \ldots, p_k$. Define A' to be an A-similar event as $A' = \{A_1, \ldots, A_k, A_{k+1}\}$ with probabilities $p_1, \ldots, p_k, p_{k+1} = 0$.

- If two sets A and B are independent $\mathbb{H}(A \times B) = \mathbb{H}(A) + \mathbb{H}(B)$.
- $\mathbb{H}(A) = \mathbb{H}(A')$.

The only function that satisfies the above two properties is $\mathbb{H}(A) = \lambda \sum_i p_i \log p_i$.
Why this result is important?

More about entropy

A_1	A_2
0.1	0.9
0.49	0.51
0.69	0.325

Dimension estimation

- Suppose the true model $f(x \mid \boldsymbol{\theta}_K)$ is in a large space with parameters $\boldsymbol{\theta}_K = (\theta_1, \ldots, \theta_k, \ldots, \theta_K)^\top$,
- We are fitting a more parsimonious model $f(x \mid \boldsymbol{\theta}_k)$ with parameters $\boldsymbol{\theta}_k = (\theta_1, \ldots, \theta_k)^\top$. The true parameter is $\boldsymbol{\theta}_0$ of dimension $K \times 1$.

$$\mathbb{KL}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_k) = \mathbb{KL}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta}) = \frac{1}{2}\Delta\boldsymbol{\theta}^\top \mathbf{I} \Delta\boldsymbol{\theta}$$

Where $\mathbf{I}$ is the Fisher information.

Suppose the projection of $\boldsymbol{\theta}_0$ is $\boldsymbol{\theta}^*$. While we approximate $\mathbb{KL}$ at $\boldsymbol{\theta}_0$ we want to remain close to $\boldsymbol{\theta}_0$ in the projection, so let's use the closest projection of $\boldsymbol{\theta}_0$, i.e. the MLE in the lower dimension $\hat{\theta}_k$.

$$
\begin{aligned}
\mathbb{KL}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_k) &\approx (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_k)^\top \mathbf{I}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_k) \\
&\approx (\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*)^\top \mathbf{I}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*) \\
&\quad + (\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_k)^\top \mathbf{I}(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_k)
\end{aligned}
$$

$$
\begin{aligned}
2n\mathbb{E}\{\mathbb{KL}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_k)\} &= n(\theta_0 - \boldsymbol{\theta}^*)^\top \mathbf{I}(\theta_0 - \boldsymbol{\theta}^*) \\
&\quad + \mathbb{E}\{n(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_k)^\top \mathbf{I}(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_k)\} \\
&= \{-2\log L(\hat{\boldsymbol{\theta}}_k) + 2k\} \\
&\quad + \{2\log L(\hat{\boldsymbol{\theta}}_K) - K\}.
\end{aligned}
$$

Considerations

- Data are i.i.d.
- $\boldsymbol{\theta} \in \mathbb{R}^{\mathrm{K}}$
- $\hat{\boldsymbol{\theta}}_k$ converges with standard rate $o_p(n^{-\frac{1}{2}})$ to $\boldsymbol{\theta}^*$
- Estimation is maximum likelihood
- k is close to K
- Local alternative asymptotic conditions hold
- $f(\mathbf{x} \mid \boldsymbol{\theta})$ is smooth with respect to $\boldsymbol{\theta}$
- Comparing models must be nested with respect to a big model of dimension K.
- Is inconsistent and tends to overfits asymptotically.

$$\begin{aligned} \mathrm{AIC} &= -2\log\text{likelihood} + 2k \\ \mathrm{TIC} &= ? \\ \mathrm{BIC} &= -2\log\text{likelihood} + \log nk \\ \mathrm{DIC} &= ? \end{aligned}$$

- Takeuchi Information Criterion (TIC): think about wrong parametric models
- Deviance Information Criterion (DIC): think about Bayesian hierarchical models

BIC

For model M with parameter vector $\boldsymbol{\theta}$ of dimension $k \times 1$, the *evidence* principle says that the data supports the model that brings more predictive power

$$f(x \mid M) = \int f(x \mid M, \boldsymbol{\theta}) f(\boldsymbol{\theta} \mid M) d\boldsymbol{\theta}$$

If $\boldsymbol{\theta}$ converges with $o_p(n^{-\frac{1}{2}})$, if one supposes $f(\boldsymbol{\theta} \mid M) = \mathrm{cst}$, the Laplace approximation gives

$$-2 \log f(\mathbf{x} \mid M) \approx -2 \log f(x \mid \hat{\boldsymbol{\theta}}, M) + k \log n$$

Summary

- BIC is a consistent model selection: $\mathbf{P}(\hat{M}_n = M) = 1$ as long as $M \in \{\mathcal{M}_n\}$ asymptotically
- Use BIC for model selection and this is equivalent to penalization with $||\boldsymbol{\beta}||_0$
- AIC tends to overfit

Leave-one-out $=$ Jackknife

$$E = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}^{(-i)})^2$$

if $\mathbf{y} = \mathbf{H}y$

$$E = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_{ii}}\right)^2$$

Where h_{ii} is the diagonal element of $\mathbf{H}$

Connections

- Put each data point into n bins.
- k-fold cross-validation: Put data into k bins
- Generalized cross validation
  $h_{ii} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{1}{n}\mathrm{tr}(H)$

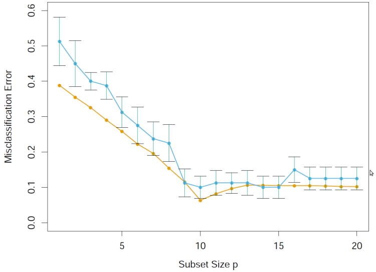
Idea

1	2	3	4	5
Train	Train	Validation	Train	Train

$$\mathrm{CV}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}^{-\kappa(i)}(\mathbf{x}_i))$$

- In regression $L(y,\hat{y})$ is the euclidean norm $(y - \hat{}y)^2$
- In classification $L(y,\hat{y}) = y \log \hat{y}$ is the cross entropy.
- Cross entropy is the multinomial negative log likelihood.

In practice



- Implement cross-validation B times: $\hat{E}_b = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}^{(b)})^2$
- $\bar{E} \pm 1.96\sqrt{\hat{\mathbb{V}}(\bar{E})} = \bar{E} \pm 1.96\frac{\hat{\sigma}_E}{\sqrt{B}}$

Cross-validation and AIC

Take $\frac{1}{(1-x)^2} \approx 1 + 2x$ and use $x = \mathrm{tr}\left(\frac{\mathbf{H}}{n}\right) = \frac{p}{n}$

$$
\begin{aligned}
E &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - \mathrm{tr}\left(\frac{\mathbf{H}}{n}\right)}\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \frac{1}{\left\{1 - \mathrm{tr}\left(\frac{\mathbf{H}}{n}\right)\right\}^2} \\
&\approx \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \left\{1 + 2\mathrm{tr}\left(\frac{\mathbf{H}}{n}\right)\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \frac{2p}{n}\hat{\sigma}^2 \\
&= \frac{\hat{\sigma}^2}{n}\left\{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2p\right\} = \frac{\hat{\sigma}^2}{n}\mathrm{AIC}
\end{aligned}
$$

Degrees of freedom

$$
\begin{aligned}
\sum_{i=1}^{n} \mathrm{cov}(y_i, \hat{y}_i) &= \mathrm{tr}\{\mathrm{cov}(\mathbf{y}, \hat{\mathbf{y}})\} \\
&= \mathrm{tr}(\mathbf{H})\mathbb{V}(\mathbf{y}) \\
&= \mathrm{tr}(\mathbf{H})\sigma^2 \\
&= p\sigma^2
\end{aligned}
$$

Regression degrees of freedom

$$
\frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{cov}(y_i, \hat{y}_i)
$$

Ridge regression

$$\mathbf{H}_\lambda = \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top$$

- $\mathrm{tr}(\mathbf{H}_\lambda)$ reflects regression degrees of freedom, depending on λ ranges from p to 0
- if β_0 is not penalized ranges from p to 1

univariate function approximation

Suppose approximation of a good univariate function over a set of observed $(x_i, y_i), i = 1, \ldots, n$.

$$y_i = f(x_i) + \varepsilon_i \approx \sum_j \beta_j b_j(x_i)$$

- polynomial base $x \in [-1, 1]$, $b_j(x_i) = x_i^j$
- Fourier base $x \in [-\pi, \pi]$,

$$y_i \approx \sum_{j=1}^{k} \beta^{(1)} \sin\left(\frac{2\pi j}{k}\right) + \beta^{(2)} \cos\left(\frac{2\pi j}{k}\right)$$

- Wavelet base of resolution k, $x \in [0, 2\pi]$

$$y_i \approx \sum_{j=1}^{2^k-1} \beta_j^{(k)} b_j^{(k)}(x_i)$$