محاسبات آماری پیشرفته
ترم اول سال تحصیلی ۹۳
جلسه یازدهم: جک‌نایف و اعتبارسنجی متقابل

حسین باغیشنی

دانشگاه شاهرود

۱۴ آذر ۱۳۹۳

جک‌نایف: یک روش بازنمونه‌گیری دیگر

روش جک‌نایف مانند خودگردان‌سازی، یک روش بازنمونه‌گیری است. در واقع به صورت ناپارامتری عمل می‌کند.

این روش خیلی زودتر از خودگردان‌سازی و بعضی دیگر از روش‌های بازنمونه‌گیری معرفی شد.

در واقع این روش در سال ۱۹۴۹ توسط کوینولا برای برآورد اریبی و سپس توسط توکی (۱۹۵۸) برای برآورد خطای استاندارد، پیشنهاد شد.

روش جک‌نایف شبیه به نوعی اعتبارسنجی متقابل LOO $(Leave\ One\ Out)$ عمل می‌کند که در مورد آن صحبت خواهیم کرد.

فرض کنید $x = (x_1, \ldots, x_n)$ نمونه مشاهده شده باشد. iامین نمونه جک‌نایف را با $x_{(i)}$ نشان می‌دهیم، به طوری که همان نمونه x است با این تفاوت که iامین مشاهده نمونه حذف شده باشد. یعنی:

$$x_{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n).$$

برآورد اریبی به روش جک‌نایف

فرض کنید تابعی مورد نظر (برای استنباط) $\theta = t(F)$ باشد.

اگر $\hat{\theta} = T_n(x)$، آنگاه برآورد iامین جک‌نایف، برای $i = ۱, \ldots, n$، عبارتست از $\hat{\theta}_{(i)} = T_{n-۱}(x_{(i)})$.

برآوردگر $\hat{\theta}$ همان برآوردگر جایگذاری‌شده بر اساس توزیع تجربی، یعنی $\hat{\theta} = t(F_n)$، است.

تعریف: یک برآوردگر جایگذاری مانند $\hat{\theta}$ را هموار گویند، هرگاه تغییرات کوچک در داده‌ها، تغییرات ناچیزی در $\hat{\theta}$ ایجاد کند.

اگر $\hat{\theta}$ یک برآوردگر هموار باشد، تعریف می‌کنیم $\hat{\theta}_{(i)} = t(F_{n-۱}(x_{(i)}))$ و برآورد اریبی جک‌نایف عبارتست از:

$$\hat{b}_{jack} = (n-۱)(\bar{\theta}_{(\cdot)} - \hat{\theta}),$$

که در آن

$$\bar{\theta}_{(\cdot)} = \frac{۱}{n}\sum_{i=۱}^{n} \hat{\theta}_{(i)}.$$

برآورد اریبی

اگر $\hat{\theta}$ نااریب باشد، آن‌گاه

$$\mathbb{E}(\bar{\theta}_{(\cdot)}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\hat{\theta}_{(i)}) = \theta.$$

اما معمولا برای برآوردگرها داریم:

$$\mathbb{E}(\hat{\theta}) = \theta + \frac{a}{n} + \frac{b}{n^2} + O(n^{-3}).$$

در نتیجه

$$\mathbb{E}(\bar{\theta}_{(\cdot)} - \hat{\theta}) = \frac{a}{n(n-1)} + O(n^{-3}).$$

بنابراین برآوردگر جک‌نایف با اریبی تصحیح‌شده، $\hat{\theta}_{jack} = \hat{\theta} - b_{jack}$، یک برآوردگر نااریب برای θ، تا مرتبه ۲، است (چرا؟)

برآورد اریبی: واریانس نمونه

برای درک بهتر روابط قبلی، فرض کنید تابعی θ واریانس جامعه باشد.

بنابراین برآوردگر جایگذاری واریانس بر حسب نمونه‌ای به حجم n عبارتست از $\hat{\theta} = ۱/n \sum_{i=۱}^{n} (x_i - \bar{x})^۲$.

این برآوردگر اریب است:

$$b(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \sigma^۲) = \frac{n-۱}{n}\sigma^۲ - \sigma^۲ = -\frac{\sigma^۲}{n}.$$

در نتیجه اریبی در هر نمونه جک‌نایف $-\frac{\sigma^۲}{n-۱}$ است. پس برای هر $i = ۱, \ldots, n$ داریم:

$$\begin{aligned}\mathbb{E}(\hat{\theta}_{(i)} - \hat{\theta}) &= \mathbb{E}(\hat{\theta}_{(i)} - \theta) - \mathbb{E}(\hat{\theta} - \theta) = b(\hat{\theta}_{(i)}) - b(\hat{\theta}) \\ &= -\frac{\sigma^2}{n-1} - (-\frac{\sigma^2}{n}) = -\frac{\sigma^2}{n(n-1)} = \frac{b(\hat{\theta})}{n-1}.\end{aligned}$$

در این‌جا برآورد با اریبی تصحیح‌شده همان برآوردگر نااریب واریانس است: $\hat{\theta}_{jack} = S^۲ = \frac{۱}{n-۱} \sum_{i=۱}^{n} (x_i - \bar{x})^۲$.

مثال: داده‌های نوع پانسمان

در مبحث روش خودگردان‌سازی، مثالی برای داده‌های پانسمان بهداشتی آورده شد. برای آن مثال، می‌خواهیم با روش جک‌نایف، اریبی را برآورد کنیم.

```
data(patch, package = "bootstrap")
n <- nrow(patch)
y <- patch$y
z <- patch$z
theta.hat <- mean(y) / mean(z)
> print (theta.hat)
[1] -0.0713061
# compute the jackknife replicates, leave-one-out estimates
theta.jack <- numeric(n)
for (i in 1:n){
  theta.jack[i] <- mean(y[-i]) / mean(z[-i])
}
  bias <- (n - 1) * (mean(theta.jack) - theta.hat)
> print(bias) # jackknife estimate of bias
[1] 0.008002488
```

شبه‌مقادیر جک‌نایف

راه دیگری برای معرفی برآوردگر جک‌نایف، بر حسب شبه‌مقادیر جک‌نایف است:

$$\tilde{\theta}_i = n\hat{\theta} - (n-۱)\hat{\theta}_{(i)}, \quad i = ۱, \ldots, n.$$

با معرفی این شبه‌مقادیر، میانگین آن‌ها همان برآوردگر با اریبی تصحیح‌شده می‌باشد:

$$\frac{۱}{n}\sum_{i=۱}^{n} \tilde{\theta}_i = n\hat{\theta} - (n-۱)\bar{\theta}_{(\cdot)} = \hat{\theta} - b_{jack}.$$

ایده پشتیبان شبه‌مقادیر، این است که به ما اجازه می‌دهند برآوردگر با اریبی تصحیح‌شده جک‌نایف را به شکل ساده میانگین n مقدار مستقل نشان دهیم.

ویژگی‌های شبه‌مقادیر

- شبه‌مقادیر $\{\tilde{\theta}_i\}$ به طور کلی مستقل نیستند. البته برای حالت خاصی که آماره خطی است، یعنی $\hat{\theta} = n^{-1} \sum_i a(x_i)$، داریم $\tilde{\theta}_i = a(x_i)$. به عنوان مثال برای میانگین $\tilde{\theta}_i = x_i$.
- بنابراین یک ایده منطقی، در نظر گرفتن $\tilde{\theta}_i$ها به عنوان تقریب‌های خطی مشاهدات iid و میانگین گرفتن از آن‌ها به عنوان برآوردگر با اریبی تصحیح‌شده جک‌نایف $\hat{\theta}_{jack}$ است.
- با قرار دادن $\tilde{S}^2$ به عنوان واریانس نمونه شبه‌مقادیر، برآوردگری برای واریانس $\hat{\theta}$، $\sigma^2(\hat{\theta})$، به صورت زیر خواهد بود:

$$v_{jack} = \frac{\tilde{S}^2}{n}.$$

برآورد جک‌نایف خطای استاندارد

برای برآورد خطای استاندارد جک‌نایف علاوه بر فرمول مبتنی بر شبه‌مقادیر، می‌توان از عبارت زیر هم، زمانی که برآوردگر $\hat{\theta}$ هموار است، استفاده کرد:

$$\hat{se}_{jack} = \left(\frac{n-۱}{n} \sum_{i=۱}^{n} (\hat{\theta}_{(i)} - \bar{\theta}_{(\cdot)})^۲ \right)^{۱/۲}.$$

تمرین: برای محاسبه خطای استاندارد جک‌نایف، چرا از ضریب $\frac{n-۱}{n}$ در فرمول بالا استفاده شده است؟

برای مثال، از برآوردهای جک‌نایف در مثال قبلی برای داده‌های پانسمان بهداشتی بیماران، استفاده می‌کنیم:

```
se <- sqrt((n-1) *
      mean((theta.jack - mean(theta.jack))^2))
> print(se)
[1] 0.1055278
```

فواصل اطمینان جک‌نایف

شبه‌مقادیر جک‌نایف کمک می‌کند تا فواصل اطمینانی به شکل زیر بتوان ساخت:

$$\hat{\theta}_{jack} \mp t_{1-\frac{\alpha}{2};n-1} se_{jack}$$

تمرین: یک تابع کلی در R به اسم $jackknife$ بنویسید که روش جک‌نایف را اجرا کند. تابع باید دارای دو ورودی باشد: x که بیانگر داده‌هاست و $theta$ که بیانگر یک تابعی است به طوری که وقتی روی x عمل می‌کند، برآوردهایی را تولید کند. این تابع باید لیستی شامل مولفه‌های زیر را برگرداند:

- $bias$: برآورد جک‌نایف اریبی
- se: برآورد جک‌نایف خطای استاندارد
- $values$: برآوردهای جک‌نایف $\{\hat{\theta}_{(i)}\}$
- $pseudo.values$: شبه‌مقادیر جک‌نایف $\{\tilde{\theta}_i\}$
- $lower.ci$: کران پایین فاصله اطمینان جک‌نایف
- $upper.ci$: کران بالای فاصله اطمینان جک‌نایف

آیا v_{jack} برآوردگر خوبی است؟

واضح است که v_{jack} به عنوان برآوردی از $\sigma^۲(\bar{x})$، خوب است زیرا معادل است با برآوردگر (ناااریب) معمول آن.

به‌طور مشابه (با منطقی مشابه)، این برآوردگر برای هر آماره خطی نیز مناسب و خوب است.

حتی اگر g تابعی به‌طور پیوسته مشتق‌پذیر در میانگین جامعه، μ، باشد، نشان داده شده است که v_{jack} برآوردگری سازگار برای واریانس $g(\bar{x})$ است:

$$\frac{v_{jack}}{\sigma^۲(g(\bar{x}))} \xrightarrow{P} ۱$$

اما مواردی وجود دارند که v_{jack} برآوردگر مناسبی برای واریانس یک برآوردگر نیست.

به‌ویژه نشان داده شده است، زمانی که برآوردگر تابعی هموار از داده‌ها نیست، v_{jack} ضعیف عمل می‌کند.

یک مثال ساده از یک برآوردگر ناهموار، میانه است.

اجرای جک‌نایف بر روی میانه

فرض کنید نمونه $\{۱, ۲, \ldots, ۹, ۱۰\}$ مشاهده شده باشد.

برآوردهای جک‌نایف چه طور خواهند بود؟

برآوردهای جک‌نایف عبارتند از ۵ تا ۵ و ۵ تا ۶

برای هر مجموعه داده با n زوج، همیشه دو مقدار یکتا برای $\hat{\theta}_{(i)}$، هر کدام با $n/۲$ تکرار، مشاهده خواهد شد.

بنابراین به نظر نمی‌رسد استفاده از این برآوردهای جک‌نایف ایده خوبی برای برآورد واریانس میانه باشد و البته هم نیست.

نشان داده شده است v_{jack} برآوردگری ناسازگار برای $\sigma^۲(\hat{\theta})$ است:

$$\frac{v_{jack}}{\sigma^۲(\hat{\theta})} \xrightarrow{d} \left(\frac{\chi^۲_۲}{۲}\right)^۲.$$

مثال: خطای استاندارد میانه

در این مثال، خطای استاندارد میانه یک نمونه ۱۰ تایی از اعداد صحیح $\{1, \ldots, 100\}$ محاسبه می‌شود:

```
set.seed(123) # for the specific example given
n <- 10
x <- sample(1:100, size = n)
# jackknife estimate of se
M <- numeric(n)
for (i in 1:n) { # leave one out
     y <- x[-i]
     M[i] <- median(y)
}
Mbar <- mean(M)
> print(sqrt((n-1)/n * sum((M - Mbar)^2)))
[1] 1.5
# bootstrap estimate of se
Mb <- replicate(1000, expr = {
      y <- sample(x, size = n, replace = TRUE)
      median(y) })
> print(sd(Mb))
[1] 13.69387
> print(x)
 [1] 29 79 41 86 91  5 50 83 51 42
> print(M)
 [1] 51 50 51 50 50 51 51 50 50 51
> print(Mb)
```

همانطور که ملاحظه می‌شود برآورد خودگردان و جک‌نایف خیلی با هم اختلاف دارند. در واقع روش جک‌نایف به دلیل ناهموار بودن میانه، موفق نیست.

جک‌نایف dـحذفی

یک نسخه دیگر از جک‌نایف که پیشنهاد شده است، جک‌نایف dـحذفی ($delete - d\ jackknife$) نامیده می‌شود.

همانطور که از نام آن پیداست، به جای هر بار حذف یک مشاهده برای محاسبه مجموعه $\{\hat{\theta}_{(i)}\}$، هر بار d مشاهده حذف می‌شود.

این رهیافت دارای مزایایی است:
از جمله نشان داده شده است اگر d به طور مناسب انتخاب شود (صفحه ۱۹۴ کتاب را ببینید)، آنگاه برآوردگر جک‌نایف dـحذفی واریانس برای میانه سازگار خواهد بود.

البته عیبی که این روش دارد، محاسبات خیلی زمان‌بر آن است. زیرا به جای محاسبه n برآورد در جک‌نایف معمولی، تعداد $\binom{n}{d}$ برآورد dـحذفی نیاز دارد که خیلی بزرگتر از n می‌تواند باشد.

اعتبارسنجی متقابل، چیزی که هر آماردانی باید بداند

اعتبارسنجی متقابل، $Cross\ Validation$، (CV) روشی است برای ارزیابی توان (پیشگویی) یک مدل آماری

هر آماردانی می‌داند که معیارهای (آماره‌های) برازش مدل، راهنماهای خوبی برای پاسخ به این سوال که یک مدل با چه دقتی پیشگویی خواهد کرد، نیستند.

به عنوان مثال، R^2 بزرگ برای یک مدل لزوماً به معنی خوب بودن آن مدل نیست:

- به سادگی می‌توان با افزودن پارامترهای بیشتر به یک مدل (به عبارتی بیش‌برازش مدل)، مقدار R^2 را نیز افزایش داد.
- در یک رگرسیون چندجمله‌ای، با افزودن جملات با مرتبه بالاتر می‌توان مدلی با برازش بهتر به داده‌ها به دست آورد.

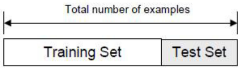
اما پیشگوییهای مدل بر روی داده‌های جدید در یک مدل بیش‌برازش‌شده بدتر خواهد بود.

بیش‌برازش = عملکرد ضعیف در پیشگویی

مجموعه‌های آموزشی و آزمون

یک روش برای اندازه‌گیری قدرت پیشگویی یک مدل، آزمودن آن بر روی مجموعه‌ای از داده‌هاست که در برازش مدل مورد استفاده قرار نگرفته باشند.

متخصصان یادگیری ماشینی و داده‌کاوها به چنین مجموعه‌ای، مجموعه آزمون و به مجموعه داده‌ای که برای برازش مدل استفاده شده است، مجموعه آموزشی می‌گویند.



به عنوان مثال، دقت پیشگویی یک مدل را می‌توان با MSE آن که بر روی مجموعه آزمون محاسبه شده است، اندازه‌گیری کرد.

به‌طور کلی، از آن‌جا که داده‌های آزمون در برازش مدل استفاده نشده‌اند، این MSE نسبت به MSE محاسبه‌شده از مجموعه آموزشی، بزرگتر خواهد بود.

مجموعه‌های آموزشی و آزمون: عیب‌ها

استفاده از این رهیافت دو عیب اساسی دارد:

۱. در مواردی که حجم داده‌ها کم است، ممکن است در نظر گرفتن بخشی از آن‌ها به عنوان مجموعه آزمون مناسب نباشد.
۲. چون این رهیافت یک آزمایش آموزش–و–آزمون تکی است، اگر تقسیم کردن داده‌ها به دو مجموعه آموزش و آزمون به‌طور مناسب انجام نشده باشد، استنباط‌ها می‌توانند گمراه‌کننده باشند.

این محدودیت‌ها با در نظر گرفتن صورت‌های پیچیده‌تری از مجموعه‌های آموزشی و آزمون مرتفع می‌شوند:

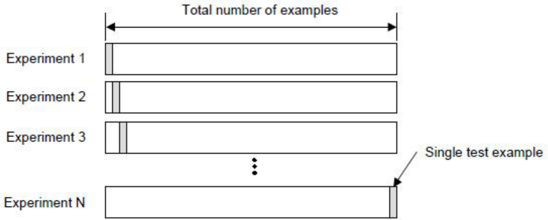
- Leave-one-out cross-validation (LOOCV)
- Leave-k-out cross-validation
- K-Fold cross-validation
- ...

$LOOCV$

فرض کنید N مشاهده مستقل $y_1, \ldots, y_N$ را داریم.

در روش $LOOCV$، اندازه‌های دقت به صورت زیر به دست می‌آیند:

1. فرض کنید مشاهده iام مجموعه آزمون را تشکیل دهد و مدل را با بقیه مشاهدات برازش دهیم. سپس خطای $e_i^* = y_i - \hat{y}_i$ را برای مشاهده حذف‌شده محاسبه می‌کنیم. به این خطا گاهی خطای پیشگویی نیز می‌گویند.



2. مرحله ۱ را برای $i = 1, \ldots, N$ تکرار می‌کنیم.

3. MSE را برحسب $e_1^*, \ldots, e_N^*$ محاسبه می‌کنیم که به آن کمترین توان‌های دوم خطای پیشگویی، $MSPE$، می‌گویند.

$LOOCV$: ادامه

این روش، استفاده خیلی کاراتری از داده‌های موجود است. زیرا در هر مرحله تنها یک مشاهده حذف می‌شود.

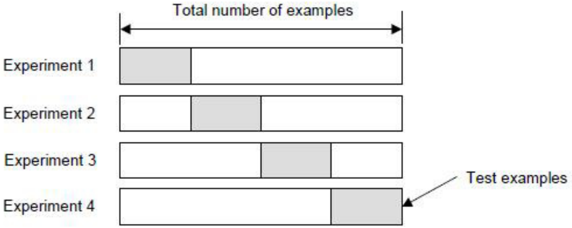
البته این روش، جز برای مدل‌های خطی، می‌تواند خیلی زمان‌بر باشد.

معیارهای دیگری مانند MAE به‌طور مشابه قابل محاسبه‌اند. یک معیار مرتبط آماره $PRESS$ است که برابر $MSPE \times N$ می‌باشد.

اعتبارسنجی متقابل با حذف هر بار k مشاهده

یکی از صورت‌های اعتبارسنجی متقابل، شامل تشکیل مجموعه‌های آزمون با حجم k مشاهده در هر مرحله است.

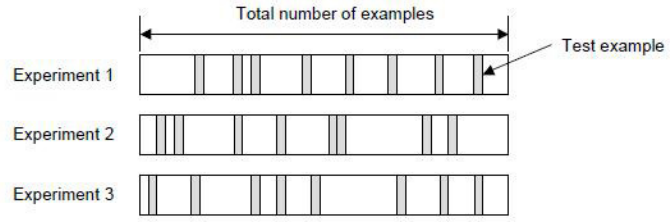
یکی از مزیت‌های این رهیافت، محاسبات کمتر آن است. اما در مورد نحوه انتخاب k باید دقت و توجه کافی داشت.



یک انتخاب معمول $k = ۱۰$ است.

اعتبارسنجی متقابل با بازنمونه‌گیری تصادفی

یکی دیگر از روش‌های اعتبارسنجی متقابل، اعتبارسنجی متقابل k $fold$ است که در آن نمونه اصلی به طور تصادفی به k زیرنمونه تقسیم می‌شود و در هر مرحله، یکی از آن‌ها به عنوان مجموعه آزمون در نظر گرفته می‌شود.



یک نسخه متداول دیگر $bootstrap$ + ۰.۶۳۲ است که توسط افرون و تیبشیرانی (۱۹۹۷) معرفی شد و ویژگی‌های بهتری دارد، اما اجرای آن پیچیده است.

Efron, B. & Tibshirani, R. (1997), *Improvements on Cross-Validation: The .632+ Bootstrap Method*, JASA, 92, 548-560

کاربردها

به‌طور کلی با می‌نیمم کردن $MSPE$ (یا آماره‌های مشابه)، می‌توان مدل آماری با قدرت پیشگویی بهتر را انتخاب کرد.

از اعتبارسنجی متقابل در قسمت‌های مختلفی از استنباط آماری، به‌ویژه انتخاب مدل، استفاده می‌شود:

- سنجش نیکویی برازش مدل‌ها و انتخاب بهترین مدل
- ارزیابی پایداری برآوردهای پارامترها
- سنجش دقت رده‌بندی الگوریتم‌های رده‌بندی
- انتخاب پارامترهای میزان‌ساز، $tuning\ parameters$، مدل‌های آماری مختلف مانند: درجه آزادی یک هموارساز ناپارامتری، پارامتر k در روش kـ نزدیکترین همسایگی، kNN، پارامترهای هسته در روش SVM.

به طور عملی، انتخاب مدل بر اساس CV خیلی بهتر از انتخاب مدل مبتنی بر آزمون‌های آماری است و تقریباً یک اندازه ناریب از MSE واقعی برای داده‌های جدید ارایه می‌کند.

مثال

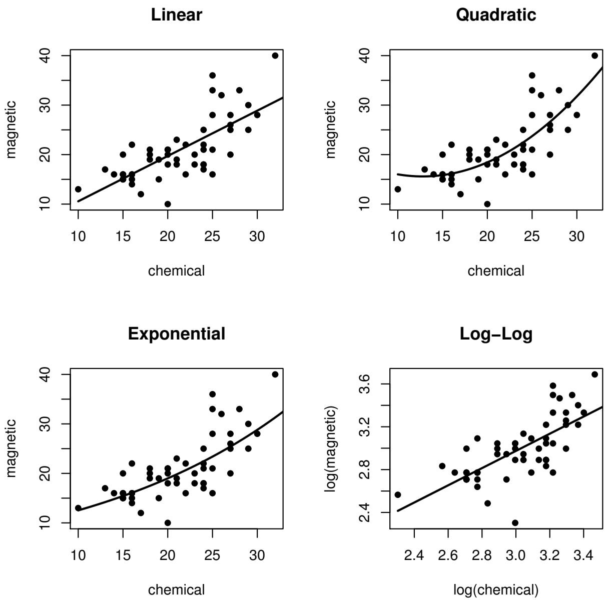
مجموعه داده $ironslag$ که در بسته $DAAG$ موجودند، شامل ۵۳ اندازه‌گیری میزان آهن به دو روش شیمیایی و مغناطیسی است.

با توجه به شکل پراکنش داده‌ها، این احساس وجود دارد که ممکن است رابطه بین این دو متغیر، خطی نباشد.

برای این داده‌ها سه رابطه دیگر: درجه دو، نمایی و لگاریتمی در نظر می‌گیریم و مدل‌های مربوط به آن‌ها را به صورت زیر معرفی می‌کنیم:

$$\begin{aligned} Linear: \quad Y &= \beta_0 + \beta_1 X + \epsilon \\ Quadratic: \quad Y &= \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon \\ Exponential: \quad \log(Y) &= \beta_0 + \beta_1 X + \epsilon \\ Logarithmic: \quad \log(Y) &= \beta_0 + \beta_1 \log(X) + \epsilon \end{aligned}$$

روش‌های متفاوتی برای انتخاب بهترین مدل، بر حسب هدف، وجود دارند. در این مثال بر روش مبتنی بر خطای پیشگویی که با CV قابل برآورد است، متمرکز می‌شویم.


Linear
Quadratic
Exponential
Log−Log
magnetic
chemical
log(magnetic)
log(chemical)
10
15
20
25
30
40
2.4
2.6
2.8
3.0
3.2
3.4
3.6

محاسبه خطای پیشگویی

بر اساس $LOOCV$ خطای پیشگویی را می‌توان به صورت زیر برآورد کرد:

۱. برای $i = ۱, \ldots, n$ مشاهده (x_i, y_i) را مجموعه آزمون در نظر بگیرید و از بقیه داده‌ها برای برازش مدل استفاده کنید:
   الف) مدل را با $n - ۱$ مشاهده (x_j, y_j)، که در آن $j \neq i$، در مجموعه آموزشی برازش دهید.
   ب) مقادیر پیشگویی $\hat{y}_i$ را برای مجموعه آزمون به دست آورید. مثلا برای مدل خطی به صورت $\hat{y}_i = \hat{\beta}_۰ + \hat{\beta}_۱ x_i$ خواهد بود.
   ج) خطای پیشگویی $e_i^* = y_i - \hat{y}_i$ را محاسبه کنید.

۲. MSE خطاهای پیشگویی را به صورت $MSPE = \frac{۱}{n} \sum_{i=۱}^{n} e_i^{*۲}$ محاسبه کنید.

مدلی که دارای کمترین $MSPE$ باشد، به عنوان بهترین مدل انتخاب می‌شود.

انتخاب بهترین مدل

```
n <- length(magnetic)   #in DAAG ironslag
e1 <- e2 <- e3 <- e4 <- numeric(n)
# fit models on leave-one-out samples
for (k in 1:n) {
     y <- magnetic[-k]
     x <- chemical[-k]
##
     J1 <- lm(y ~ x)
     yhat1 <- J1$coef[1] + J1$coef[2] * chemical[k]
     e1[k] <- magnetic[k] - yhat1
##
     J2 <- lm(y ~ x + I(x^2))
     yhat2 <- J2$coef[1] + J2$coef[2] * chemical[k] +
              J2$coef[3] * chemical[k]^2
     e2[k] <- magnetic[k] - yhat2
##
     J3 <- lm(log(y) ~ x)
     logyhat3 <- J3$coef[1] + J3$coef[2] * chemical[k]
     yhat3 <- exp(logyhat3)
     e3[k] <- magnetic[k] - yhat3
##
     J4 <- lm(log(y) ~ log(x))
     logyhat4 <- J4$coef[1] + J4$coef[2] * log(chemical[k])
     yhat4 <- exp(logyhat4)
     e4[k] <- magnetic[k] - yhat4
}
> c(mean(e1^2), mean(e2^2), mean(e3^2), mean(e4^2))
[1] 19.55644 17.85248 18.44188 20.45424
```

آیا همیشه می‌توان از CV استفاده کرد؟

دقت کنید که اعتبارسنجی متقابل همیشه قابل به کارگیری نیست.

به عنوان مثال، در یک مدل رگرسیونی، اگر دو یا بیشتر از دو مشاهده با مقادیر دقیقاً یکسان برای همه متغیرهای تبیینی و متغیر پاسخ y وجود داشته باشد، آنگاه خارج کردن یک مشاهده ($LOOCV$) موثر نخواهد بود.

برآورد سازگار مدل با CV

یک معیار انتخاب مدل را سازگار گویند، هرگاه اگر، در رده‌ای از مدل‌ها، مدل درست وجود داشته باشد، بتواند با افزایش n آن را شناسایی کند.

در یک مقاله معروف، شائو (۱۹۹۷) نشان داد $LOOCV$ به برآوردی سازگار منتهی نمی‌شود.

Shao, J. (1997), *An Asymptotic Theory for Linear Model Selection*, Statistica Sinica, 7, 221-264

به عبارت دیگر $LOOCV$، معیار سازگاری برای انتخاب مدل نیست. یعنی $LOOCV$ همواره قادر به پیدا کردن مدل درست نیست (البته به گفته مشهور باکس باید توجه داشت):

همه مدل‌ها نادرستند، اما برخی از آن‌ها مفیدند

به عبارتی، پذیرفتن وجود مدل درست در رده مدل‌های مورد نظر، چندان واقع‌بینانه نیست!!

برآورد سازگار مدل با CV : ادامه

در مقابل، انواع مشخصی از اعتبارسنجی متقابل با حذف k مشاهده (به‌طور دقیق‌تر زمانی که k همراه با n افزایش می‌یابد)، سازگار خواهند بود.

در عمل، یک مشکل این معیار آن است که تغییر کوچک در داده‌ها می‌تواند باعث تغییر زیاد در مدل انتخاب شده شود.

محققین مختلفی دریافته‌اند که اعتبارسنجی متقابل $k\ fold$، از این منظر، دارای عملکرد بهتری است.

اعتبارسنجی متقابل برای مدل‌های خطی

به‌طور کلی CV می‌تواند از نظر محاسباتی بسیار زمان‌بر و پرهزینه باشد.

اما برای مدل‌های خطی، محاسبه $LOOCV$ بسیار سریع و ساده است.

مدل خطی را در نظر بگیرید:

$$Y = X\beta + \epsilon.$$

می‌دانیم

$$\hat{\beta} = (X^{\prime}X)^{-1}X^{\prime}Y$$

و مقادیر برازش‌شده به صورت زیر محاسبه می‌شوند:

$$\hat{Y} = X\hat{\beta} = X(X^{\prime}X)^{-1}X^{\prime}Y = HY$$

چون برای محاسبه $\hat{Y}$ از H استفاده می‌شود، به آن ماتریس hat می‌گویند.

اعتبارسنجی متقابل برای مدل‌های خطی: ادامه

اگر مقادیر روی قطر اصلی H را با $h_۱, \ldots, h_n$ نشان دهیم، آماره اعتبارسنجی متقابل (همان MSE باقی‌مانده‌های پیشگویی e_i^*) به صورت زیر قابل محاسبه است:

$$MSPE = \frac{۱}{n}\sum_{i=۱}^{n}\left[\frac{e_i}{۱-h_i}\right]^۲,$$

که در آن، e_i باقی‌مانده‌های حاصل از برازش مدل بر روی همه n مشاهده است.

بنابراین برای مدل خطی لازم نیست n مدل جدا برازش داده شود.

این نتیجه بسیار جالب، این امکان را میسر می‌سازد که برای محاسبه $MSPE$ تنها یک بار مدل بر روی کل داده‌ها برازش داده شود.

رابطه با سایر معیارها

آماره‌های CV و معیارهای مرتبط با آن به شدت در آمار طرفدار دارند و مورد استفاده قرار می‌گیرند.

وجود رابطه سایر معیارهای انتخاب مدل با آماره‌های CV، به‌طور واضح، نشان داده نشده است. در این‌جا به چند مورد شناخته‌شده اشاره می‌کنیم:

- جک‌نایف: همان طور که در اسلایدهای قبلی مطرح شد، برآوردگر جک‌نایف مشابه $LOOCV$ محاسبه می‌شود، با این تفاوت که به جای محاسبه باقی‌مانده‌های e_i^* در هر تکرار، تابعی مورد نظر θ محاسبه می‌شود.
- AIC: معیار اطلاع آکاییک به صورت $AIC = -۲ \log \mathcal{L} + ۲p$ تعریف می‌شود که در آن، $\mathcal{L}$ مقدار ماکسیمم تابع درستنمایی و p تعداد پارامترهای آزاد مدل است. به‌طور مجانبی، می‌نیمم کردن AIC معادل می‌نیمم کردن آماره CV است. این نتیجه برای هر مدلی، نه فقط مدل خطی، برقرار است. این ویژگی باعث شده است زمانی که هدف پیشگویی است، استفاده از AIC توصیه شود.

رابطه با سایر معیارها: ادامه

- BIC: معیار اطلاع بیزی به صورت $BIC = -۲ \log \mathcal{L} + p \log(n)$ تعریف می‌شود. به دلیل جریمه سنگین‌تر این معیار نسبت به AIC، مدلی که توسط BIC انتخاب می‌شود یا همان مدل منتخب بر اساس AIC است یا مدلی با تعداد پارامتر کمتر. شائو (۱۹۹۷) نشان داد، برای مدل‌های خطی، به‌طور مجانبی می‌نیمم کردن BIC معادل اعتبارسنجی متقابل با حذف ν مشاهده است، به‌طوری که

$$\nu = n[۱ - ۱/(\log(n) - ۱)].$$

یک منبع جامع و عالی برای روش‌های اعتبارسنجی متقابل، آرلوت و سلیسه (۲۰۱۰) است.

Arlot, S. & Celisse, A. (2010), *A Survey of Cross-Validation Procedures for Model Selection*, Statistics Surveys, 4, 40-79

اعتبارسنجی متقابل برای سری‌های زمانی

وقتی که داده‌ها مستقل نیستند، اجرای CV خیلی سختتر می‌شود، زیرا کنار گذاشتن یک مشاهده تمام اطلاعات مرتبط با آن را، به دلیل وابستگی با سایر مشاهدات، حذف نمی‌کند.

برای حالت خاصی که وابستگی بین داده‌ها به دلیل ماهیت وابسته به زمان بودن آن‌هاست، (یعنی در سری‌های زمانی)، یک آماره CV به صورت زیر به دست می‌آید:

۱. مدل را به داده‌های $y_۱, \ldots, y_t$ برازش داده و $\hat{y}_{t+۱}$ را مقدار پیش‌بینی برای مشاهده بعدی قرار می‌دهیم. سپس خطای $e^*_{t+۱} = y_{t+۱} - \hat{y}_{t+۱}$ را به دست می‌آوریم.
۲. مرحله ۱ را برای $t = m, \ldots, n - ۱$ تکرار می‌کنیم، که در آن m می‌نیمم تعداد مشاهداتی است که برای برازش مدل نیاز داریم.
۳. مقدار MSE را برای $e^*_n, \ldots, e^*_{m+۱}$ محاسبه می‌کنیم.

اعتبارسنجی متقابل برای داده‌های وابسته با ساختارهای وابستگی پیچیده، مانند مدل‌های آمیخته یا فضایی، می‌تواند خیلی مشکل و پرهزینه باشد.

اجرای CV در R

در R بسته‌ها و توابع متفاوتی وجود دارند که در رده‌های مشخصی از مدل‌های آماری، اعتبارسنجی متقابل را اجرا می‌کنند. در این جا به سه تا از آن‌ها اشاره می‌کنیم:

- تابع $validate$ در بسته $Design$ برای مدل‌های خطی و لجستیک، روش CV را اجرا می‌کند. این تابع، اعتبارسنجی مبتنی بر روش خودگردان‌سازی و همچنین $bootstrap + ۰٫۶۳۲$ را نیز به عنوان انتخاب‌هایی فراهم آورده است.
- بسته $DAAG$: این بسته دارای سه تابع $cv.lm$، $CVlm$ و $CVbinary$ است که روش اعتبارسنجی متقابل با بازنمونه‌گیری تصادفی را در مدل‌های به ترتیب رگرسیون ساده، رگرسیون چندگانه و لجستیک، اجرا می‌کند.
- بسته $boot$: تابع $cv.glm$ روش اعتبارسنجی متقابل با بازنمونه‌گیری تصادفی k تایی را برای مدل‌های خطی تعمیم‌یافته (GLM) از جمله دوجمله‌ای، گاوسی، پواسون، گاما و غیره، اجرا می‌کند. اگر k برابر تعداد مشاهدات مشخص شود، آن‌گاه روش $LOOCV$ خواهد بود که پیش‌فرض تابع نیز همین است.