Model Selection

Содержание

- Модель
$$\boldsymbol{\eta} = \mathbf{f}(\boldsymbol{\xi}) + \boldsymbol{\epsilon}, \quad \operatorname{cov} \boldsymbol{\epsilon} = \sigma^2 \mathbf{I},\ \epsilon_i \perp\!\!\!\perp \epsilon_j$$

- Функция потерь
$$L(f(\boldsymbol{\xi}), \eta) = |f(\boldsymbol{\xi}) - \eta|^d, \quad d \in \{1, 2\}.$$

- Пусть модель $\hat{f}$ обучена на тренировочном множестве $\mathcal{T}$; тогда хотелось бы знать её generalization (test, prediction) error
$$\mathrm{Err}_{\mathcal{T}} = \mathsf{E}\left[L(\hat{f}(\boldsymbol{\xi}), \eta) \mid \mathcal{T}\right]$$
т.е. ожидание функции потерь на генеральной совокупности или
$$\mathrm{Err} = \mathsf{E}\left[\mathrm{Err}_{\mathcal{T}}\right] = \mathsf{E}\left[L(\hat{f}(\boldsymbol{\xi}), \eta)\right]$$
т.е. ожидание еще и по всем тренировочным выборкам. Тогда из всех возможных моделей могли бы выбрать $\hat{f}$, минимизирующую $\mathrm{Err}_{\mathcal{T}}$ (или Err?)

- Справедливо для квадратичной L,
$$\begin{aligned}
\mathrm{Err} &= \mathsf{E}\left[L(\hat{f}(\boldsymbol{\xi}), \eta)\right] = \mathsf{E}\left[(\hat{f}(\boldsymbol{\xi}) - \eta)^2\right] = \mathsf{E}\left[\left(\left(\hat{f}(\boldsymbol{\xi}) - f(\boldsymbol{\xi})\right) + \epsilon\right)^2\right] \\
&= \mathsf{E}\left[\left(\hat{f}(\xi) - f(\xi)\right)^2 + 2\epsilon\left(\hat{f}(\xi) - f(\xi)\right) + \epsilon^2\right] \\
&= \mathsf{E}\left[\left(\hat{f}(\xi) - f(\xi)\right)^2\right] + \sigma^2.
\end{aligned}$$
Видно, что $\mathsf{E}\left[\left(\hat{f}(\xi) - f(\xi)\right)^2\right]$ потенциально можно свести к 0, тогда как фиксированную σ^2 — нет. Значит, нулевой ошибки для модели $\hat{f}$ добиться физически невозможно (её минимальной величиной и будет σ^2).

- Для $\mathbf{x}_0 \notin \mathcal{T}$, и квадратичной функции потерь справедливо[1] разложение
$$\mathrm{Err}(\mathbf{x}_0) = \mathsf{E}\left[L(\hat{f}(\mathbf{x}_0), \eta) \mid \boldsymbol{\xi} = \mathbf{x}_0\right] = \sigma^2 + \underbrace{(\mathsf{E}\hat{f}(\mathbf{x}_0) - f(\mathbf{x}_0))^2}_{\mathrm{bias}^2} + \underbrace{\mathsf{E}(\hat{f}(\mathbf{x}_0) - \mathsf{E}\hat{f}(\mathbf{x}_0))^2}_{\mathsf{D}\hat{f}(\mathbf{x}_0) = \mathrm{variance}}$$

  bias показывает, насколько ожидание модели отличается от истинного значения,

  variance показывает, насколько изменится оценка $\hat{f}$, если изменится тренировочная выборка $\mathcal{T}$.

[1] http://robjhyndman.com/hyndsight/files/2015/08/2-biasvardecomp.pdf

С ростом гибкости модели уменьшаяется смещение, но растет разброс (модель начинает аппроксимировать шум ϵ) и наоборот. Стоит задача по тренировочной выборке подобрать такую гибкость модели, что $\mathrm{Err}_{\mathcal{T}}$ минимальна. Эмпирический способ это сделать — в противовес аналитическим методам вроде информационных критериев — для модели с заданной гибкости оценить $\mathrm{Err}_{\mathcal{T}}$.

- Пусть $\mathcal{S}$ — test set; величина

$$\widehat{\mathrm{Err}}_{\mathcal{T}}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}} L(\hat{f}(\mathbf{x}_i), y_i) = \operatorname*{mean}_{\mathcal{S}} L$$

  будет иметь играть роль оценки $\mathrm{Err}_{\mathcal{T}}$ (model assessment).

- Для нахождения оптимальной гибкости могли бы непосредственно посчитать training error

$$\overline{\mathrm{err}}_{\mathcal{T}} = \frac{1}{N} \sum_{i=1}^{N} L\left(\hat{f}(\mathbf{x}_i), y_i\right) = \operatorname*{mean}_{\mathcal{T}} L.$$

  Однако $\overline{\mathrm{err}}_{\mathcal{T}}$ будет убывать с увеличением гибкости модели, в то время, как $\mathrm{Err}_{\mathcal{T}}$ с какого-то момента начнет расти.

- Поэтому тренировочную выборку следует разделить на собственно тренировочную (по которой строится модель) и валидационную $\mathcal{V}$, по которой оценивать $\mathrm{Err}_{\mathcal{T}}$. Изменением гиперпараметров модели (регулирующих гибкость), минимизировать

$$\widehat{\mathrm{Err}}_{\mathcal{T}}(\mathcal{V}) = \operatorname*{mean}_{\mathcal{V}} L.$$

  Так осуществляется задача model selection.

- Если выборка мала и мало $\mathcal{V}$, следует воспользоваться K-fold cross-validation: случайным образом разбить $\mathcal{T}$ на $\mathcal{T}_1, \ldots, \mathcal{T}_K$, затем обучать модель $\hat{f}^{(-k)}$ на $\mathcal{T} \setminus \mathcal{T}_k$, а валидировать на $\mathcal{T}_k$; результаты затем усреднить:

$$\frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{|\mathcal{T}_k|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{T}_k} \left| \hat{f}^{(-k)}(\mathbf{x}_i) - y_i \right|^d \right) = \operatorname*{mean}_{\{\mathcal{T}_1, \ldots, \mathcal{T}_K\}} \left(\operatorname*{mean}_{\mathcal{T}_k} L_k \right).$$

  - При $K = N$ тренировочные выборки почти не отличаются друг от дружки, поэтому, по сравнению с другими вариантами, разброс будет большим как среднее скоррелированных величин, а смещение малым.
  - При $K \to 1$, наборот, разброс падает, смещение растет.