Cross Validation

Andrew Nobel

October, 2021

Stochastic Framework for Classification

Recall

- Observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ iid $\sim (X, Y)$
- Population, new sample $(X, Y) \in \mathcal{X} \times \{0, 1\}$ (unobserved)

How do we use observations?

- For *training*: To produce a classification rule
- For *testing*: To assess the performance of the rule we produced
- (Also for validation, to select among competing rules)

Issue: Same observations sometimes used for more than one task

Rules, Procedures, and Schemes

Recall: A *classification rule* is a map $\phi : \mathcal{X} \to \{0,1\}$

Definition: An n-sample *classification procedure* is a map

$$\phi_n : \mathcal{X} \times (\mathcal{X} \times \{0,1\})^n \to \{0,1\}$$

Given observations D_n the procedure ϕ_n yields a rule $\hat{\phi}_n(x) = \phi_n(x : D_n)$

- If D_n is random then $\hat{\phi}_n(x)$ is random
- Different data sets yield different classification rules

Definition: A *classification scheme* is a sequence $\phi_1, \phi_2, \ldots$ of procedures, one for each sample size

Example: Two Procedures, Two Datasets

- ⋆ Two n-sample procedures, e.g., $\phi_n =$ LDA and $\psi_n =$ LogReg
- ⋆ Two data sets for some task of interest, D_n^a and D_n^b

1. Apply LDA and LogReg to data D_n^a
   - $\hat{\phi}_n^a(x) = \phi_n(x : D_n^a)$ and $\hat{\psi}_n^a(x) = \psi_n(x : D_n^a)$
   - How do resulting rules differ? Is one better than the other?
2. Apply LDA to data sets D_n^a and D_n^b
   - $\hat{\phi}_n^a(x) = \phi_n(x : D_n^a)$ and $\hat{\phi}_n^b(x) = \phi_n(x : D_n^b)$
   - Does LDA produce similar rules on two data sets?

Related Issues

Stability: Does a small change in one of the data points yields a big change in the rule $\hat{\phi}_n$?

Aggregation: How can we combine different rules to get better ones?

Risk of Rules and Procedures

Risk of Rules and Procedures

Recall: A rule $\phi : \mathcal{X} \to \{0,1\}$ has *risk* $R(\phi) = \mathbb{P}(\phi(X) \neq Y)$

For a *procedure* $\phi_n : \mathcal{X} \times (\mathcal{X} \times \{0,1\})^n \to \{0,1\}$ there are two types of risk

1. **Conditional risk** $R(\hat{\phi}_n) = \mathbb{P}(\phi_n(X : D_n) \neq Y \mid D_n)$
   - Performance of rule $\hat{\phi}_n$ produced from specific data set D_n
   - $R(\hat{\phi}_n)$ is a random variable, a function of observations D_n
2. **Expected risk** $\mathbb{E}R(\hat{\phi}_n) = \mathbb{P}(\phi_n(X : D_n) \neq Y)$
   - Expected performance of procedure ϕ_n on data sets D_n
   - $\mathbb{E}R(\hat{\phi}_n)$ is a number

Risk of Rules and Procedures, cont.

⋆ Conditional risk $R(\hat{\phi}_n)$ is the performance of the *rule* $\hat{\phi}_n$

⋆ Expected risk $\mathbb{E}R(\hat{\phi}_n)$ is the expected performance of the *procedure* ϕ_n on data sets D_n

Use of risk measures

- Assessing performance of a rule or procedure
- Comparing or selecting among competing procedures
- Assessing the intrinsic difficulty of the classification problem

Estimating Risk

Problem: Risk measures depend on the unknown distribution of (X, Y)

One solution

- Replace probabilities and expectations by averages over observations
- Appeal to the law of large numbers and probability inequalities

Sample Error Rate

Definition: Given observations $D_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ the *sample error rate* or *empirical risk* of a rule $\phi : \mathcal{X} \to \{0, 1\}$ on D_n is

$$\hat{R}_n(\phi) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\phi(X_i) \neq Y_i)$$

Fact: When ϕ is fixed

1. $\mathbb{E}[\hat{R}_n(\phi)] = R(\phi)$ and $\mathrm{Var}(\hat{R}_n(\phi)) = n^{-1} R(\phi)(1 - R(\phi))$
2. $\hat{R}_n(\phi) \sim n^{-1}$ Bin$(n, R(\phi))$
3. $\hat{R}_n(\phi) \to R(\phi)$ in probability as n tends to infinity

Sample Error Rate vs Risk for Fixed Rules

Chebyshev: If ϕ is fixed, for every $t > 0$ we have

$$\mathbb{P}\left(|\hat{R}_n(\phi) - R(\phi)| \geq t\right) \;\leq\; \frac{R(\phi)(1 - R(\phi))}{n\, t^2} \;\leq\; \frac{1}{4\, n\, t^2}$$

Hoeffding: If ϕ is fixed, for every $t > 0$ we have

$$\mathbb{P}\left(|\hat{R}_n(\phi) - R(\phi)| \geq t\right) \;\leq\; 2\exp\{-2nt^2\}$$

Upshot: For a fixed rule ϕ, the sample error rate $\hat{R}_n(\phi)$ can provide a good estimate of risk $R(\phi)$ when n is moderately large

Example: Sample Size Calculation

Task: Assess risk of a rule based on iid observations D_n. Let $\delta, \epsilon > 0$. How large must n be to ensure that

$$\Pr\left(|\hat{R}_n(\phi) - R(\phi)| \geq \delta\right) \leq \epsilon$$

This says that sample error rate is close to the true risk with high probability: in ML terminology *probably almost correct* (PAC)

Solution: Consider Chebyshev and Hoeffding bounds for the probability on the left. Set the bound equal to ϵ and solve for n.

$$n_C = \frac{1}{4\,\delta^2\,\epsilon} \qquad n_H = \frac{1}{2\,\delta^2}\log\left(\frac{2}{\epsilon}\right)$$

Training and Test Sets

Training Sets and Training Error

New: Suppose rule $\hat{\phi}_n(x) = \phi_n(x : D_n)$ obtained from observations D_n

- Refer to D_n as a *training set* and $\hat{R}_n(\hat{\phi}_n)$ as *training error*

Q: Is training error $\hat{R}_n(\hat{\phi}_n)$ a good estimate of the conditional risk $R(\hat{\phi}_n)$?

A: No! Root of the problem: $\hat{\phi}_n$ and $\hat{R}_n$ based on *same* observations D_n

- In general, we expect that $\hat{R}_n(\hat{\phi}_n)$ will *underestimate* $R(\hat{\phi}_n)$
- Rule $\hat{\phi}_n$ is fit to D_n: it is likely to perform worse on another set D_n'

Example: Training error of 1-NN rules is always zero!

One Solution: Separate Training and Test Sets

1. Split iid observations $(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})$ into two disjoint groups

   - Training set $D_n = (X_1, Y_1), \ldots, (X_n, Y_n)$
   - Test set $D_m = (X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})$

   Note that training set D_n and test set D_m are independent

2. Use training set D_n to construct a classification rule $\hat{\phi}_n(x) = \phi_n(x : D_n)$

3. Assess performance of $\hat{\phi}_n$ via its average error rate on test set D_m

$$\hat{R}_m(\hat{\phi}_n) = m^{-1} \sum_{j=1}^{m} \mathbb{I}(\hat{\phi}_n(X_{n+j}) \neq Y_{n+j})$$

Training and Test Sets, cont.

Fact: Training set D_n and test set D_m are independent

1. $\mathbb{E}[\hat{R}_m(\hat{\phi}_n) \,|\, D_n] = \mathbb{P}(\hat{\phi}_n(X) \neq Y \,|\, D_n) = R(\hat{\phi}_n)$

2. For each $t > 0$,

$$\mathbb{P}\left(|\hat{R}_m(\hat{\phi}_n) - R(\hat{\phi}_n)| > t \,|\, D_n\right) \;\leq\; \exp\{-2mt^2\}$$

Downside: When data is hard to come by or expensive to obtain, splitting observations into training and test sets is a luxury, not always feasible

Cross Validation

Overview of Cross Validation

1. Split observations into k equal size groups, called "folds"

2. For each group $j = 1, \ldots, k$
   - Produce a rule from the observations *outside* group j
   - Find the error rate of the rule using the observations *inside* group j

3. Average the error rates obtained from different groups

Cross-Validation in Detail

Ingredients

- Observations $D = (X_1, Y_1), \ldots, (X_N, Y_N)$
- Number of folds $k \geq 2$
- Assume that $N = k\, m$
- Classification procedure ϕ_{N-m}

Cross Validation

1. Randomly divide D_N into k sets $D_m^{(1)}, \ldots, D_m^{(k)}$ each with m points

2. For $j = 1, \ldots, k$ do

   - Obtain rule $\hat{\phi}^j(x)$ by applying ϕ_{N-m} to training set $D_N \setminus D_m^{(j)}$

   - Let $\hat{R}^j =$ sample error rate of rule $\hat{\phi}^\ell$ on hold-out test set $D_m^{(j)}$

3. The k-fold cross validated risk estimate is the average of the sample errors

$$\hat{R}^{\text{k-CV}} := \frac{1}{k} \sum_{j=1}^{k} \hat{R}^j$$

Analysis: What is Cross Validation Estimating?

Fact: $\mathbb{E}(\hat{R}^{\text{k-CV}}) = \mathbb{E}R(\hat{\phi}_{N-m})$

- $\hat{R}^{\text{k-CV}}$ estimating expected risk rather than conditional risk
- $\hat{R}^{\text{k-CV}}$ centered at the expected risk of ϕ_{N-m}

Fact: The mean squared error $\mathbb{E}(\hat{R}^{\text{k-CV}} - \mathbb{E}R(\hat{\phi}_N))^2$ of $\hat{R}^{\text{k-CV}}$ has bias-variance decomposition

$$\text{MSE}(\hat{R}^{\text{k-CV}}) \;=\; [\,\mathbb{E}R(\hat{\phi}_{N-m}) - \mathbb{E}R(\hat{\phi}_N)\,]^2 \;+\; \mathrm{Var}(\hat{R}^{\text{k-CV}})$$

- Bias term $[\,\mathbb{E}R(\hat{\phi}_{N-m}) - \mathbb{E}R(\hat{\phi}_N)\,]^2$ usually gets smaller as k gets bigger
- Variance $\mathrm{Var}(\hat{R}^{\text{k-CV}})$ usually gets bigger as k gets bigger