Machine Learning (CSE 446): Practical Issues (continued)

Noah Smith

University of Washington
nasmith@cs.washington.edu

October 20, 2017

Estimating Performance

We do this for two reasons:

1. To select hyperparameter values (tuning)
2. To estimate a final classifier's quality on $\mathcal{D}$ (testing)

Remember that $\hat{\mathrm{A}}$, $\hat{\mathrm{P}}$, $\hat{\mathrm{R}}$, and $\hat{F}_1$ are all *estimates* of the classifier's quality under the true data distribution $\mathcal{D}$.

- Estimates are noisy!

Cross-Validation for Hyperparameter Tuning

Data: training data D, trainable classifier family $\mathcal{F}$, set of possible hyperparameter settings $\alpha^1, \ldots, \alpha^H$

Result: hyperparameter setting

partition D randomly into equal-sized *folds*, $D^1, \ldots, D^K$;

for $h \in \{1, \ldots, H\}$ **do**

  for $k \in \{1, \ldots, K\}$ **do**

    train $f^{(h,k)} \in \mathcal{F}$ on $D \setminus D^k$ with hyperparameter setting α^h;

    $\hat{\mathrm{A}}^{(h,k)} = \hat{\mathrm{A}}(f^{(h,k)})$ (or other quality score) estimated on D^k;

  end

  $\hat{\mathrm{A}}^h = \frac{1}{K} \sum_{k=1}^{K} \hat{\mathrm{A}}^{(h,k)}$;

end

return $\alpha^{\left(\mathrm{argmax}_h \hat{\mathrm{A}}^h\right)}$ (or $f \in \mathcal{F}$ trained on $\alpha^{\left(\mathrm{argmax}_h \hat{\mathrm{A}}^h\right)}$);

Algorithm 1: CrossValidateToTune

Cross-Validation for Testing

Data: data D, trainable classifier family $\mathcal{F}$
Result: accuracy estimate
partition D randomly into equal-sized *folds*, $D^1, \ldots, D^K$;
for $k \in \{1, \ldots, K\}$ **do**
  train $f^k \in \mathcal{F}$ on $D \setminus D^k$ (possibly using $\textsc{CrossValidateToTune}$ to set hyperparameters);
  $\hat{\mathrm{A}}^k = \hat{\mathrm{A}}(f^k)$ (or other quality score) estimated on D^k;
end
$\hat{\mathrm{A}} = \frac{1}{K} \sum_{k=1}^{K} \hat{\mathrm{A}}^k$;
return $\hat{\mathrm{A}}$;

Algorithm 2: $\textsc{CrossValidateToTest}$

Careful!

If you repeatedly run $\textsc{CrossValidateToTest}$ on a single dataset D, you risk overfitting to D and getting a bad estimate.

Statistical Significance

Suppose we have two classifiers, f_1 and f_2.

Statistical Significance

Suppose we have two classifiers, f_1 and f_2.

Is f_1 better? The "null hypothesis," denoted H_0, is that it isn't. But if $\hat{\mathrm{A}}_1 \gg \hat{\mathrm{A}}_2$, we are tempted to believe otherwise.

Statistical Significance

Suppose we have two classifiers, f_1 and f_2.

Is f_1 better? The "null hypothesis," denoted H_0, is that it isn't. But if $\hat{\mathrm{A}}_1 \gg \hat{\mathrm{A}}_2$, we are tempted to believe otherwise.

How much larger must $\hat{\mathrm{A}}_1$ be than $\hat{\mathrm{A}}_2$ to *reject* H_0?

Statistical Significance

Suppose we have two classifiers, f_1 and f_2.

Is f_1 better? The "null hypothesis," denoted H_0, is that it isn't. But if $\hat{\mathrm{A}}_1 \gg \hat{\mathrm{A}}_2$, we are tempted to believe otherwise.

How much larger must $\hat{\mathrm{A}}_1$ be than $\hat{\mathrm{A}}_2$ to *reject* H_0?

One view: how (im)probable is the observed difference, given $H_0 = \text{true}$?

Statistical Significance

Suppose we have two classifiers, f_1 and f_2.

Is f_1 better? The "null hypothesis," denoted H_0, is that it isn't. But if $\hat{\mathrm{A}}_1 \gg \hat{\mathrm{A}}_2$, we are tempted to believe otherwise.

How much larger must $\hat{\mathrm{A}}_1$ be than $\hat{\mathrm{A}}_2$ to *reject* H_0?

One view: how (im)probable is the observed difference, given $H_0 = \text{true}$?

Caution: statistical significance is neither necessary nor sufficient for research significance or interestingness!

A Hypothesis Test for Text Classifiers

McNemar (1947)

1. The null hypothesis: $\mathrm{A}_1 = \mathrm{A}_2$
2. Pick significance level α, an "acceptably" high probability of incorrectly rejecting H_0.
3. Calculate the test statistic, k (explained in the next slide).
4. Calculate the probability of a *more extreme* value of k, assuming H_0 is true; this is the p-value.
5. Reject the null hypothesis if the p-value is less than α.

The p-value is $p(\text{this observation} \mid H_0 \text{ is true})$, not the other way around!

McNemar's Test: Details

Assumptions: independent (test) samples and binary measurements. Count test set error patterns:

	f_1 is incorrect	f_1 is correct	
f_2 is incorrect	c_{00}	c_{10}	
f_2 is correct	c_{01}	c_{11}	$m \cdot \hat{\mathrm{A}}_2$
		$m \cdot \hat{\mathrm{A}}_1$	

If $\mathrm{A}_1 = \mathrm{A}_2$, then c_{01} and c_{10} are each distributed according to $\mathrm{Binomial}(c_{01} + c_{10}, \frac{1}{2})$.

$$\text{test statistic } k = \min\{c_{01}, c_{10}\}$$

$$p\text{-value} = \frac{1}{2^{c_{01}+c_{10}-1}} \sum_{j=0}^{k} \binom{c_{01} + c_{10}}{j}$$

Other Tests

Different tests make different assumptions.

Sometimes we calculate an interval that would be "unsurprising" under H_0 and test whether a test statistic falls in that interval (e.g., t-test and Wald test).

In many cases, there is no closed form for estimating p-values, so we use random approximations (e.g., permutation test and paired bootstrap test).

If you do lots of tests, you need to correct for that! The first thing to learn is the Bonferroni correction.

Read lots more in Daume (2017), chapter 5.7.

Bias-Variance Tradeoff

Let $\mathcal{F}$ denote the set of all possible classifiers under consideration. (E.g., all linear classifiers for the set of features we have chosen.)

Bias-Variance Tradeoff

Let $\mathcal{F}$ denote the set of all possible classifiers under consideration. (E.g., all linear classifiers for the set of features we have chosen.)

$$\epsilon(f) = \underbrace{\epsilon(f) - \min_{f^* \in \mathcal{F}} \epsilon(f^*)}_{\text{estimation error}} + \underbrace{\min_{f^* \in \mathcal{F}} \epsilon(f^*)}_{\text{approximation error}}$$

Bias-Variance Tradeoff

Let $\mathcal{F}$ denote the set of all possible classifiers under consideration. (E.g., all linear classifiers for the set of features we have chosen.)

$$\epsilon(f) = \underbrace{\epsilon(f) - \min_{f^* \in \mathcal{F}} \epsilon(f^*)}_{\text{estimation error}} + \underbrace{\min_{f^* \in \mathcal{F}} \epsilon(f^*)}_{\text{approximation error}}$$

We could maybe correct estimation error by getting more training data.

Bias-Variance Tradeoff

Let $\mathcal{F}$ denote the set of all possible classifiers under consideration. (E.g., all linear classifiers for the set of features we have chosen.)

$$\epsilon(f) = \underbrace{\epsilon(f) - \min_{f^* \in \mathcal{F}} \epsilon(f^*)}_{\text{estimation error}} + \underbrace{\min_{f^* \in \mathcal{F}} \epsilon(f^*)}_{\text{approximation error}}$$

We could maybe correct estimation error by getting more training data. More generally, we often refer to estimation error as **variance**.

Bias-Variance Tradeoff

Let $\mathcal{F}$ denote the set of all possible classifiers under consideration. (E.g., all linear classifiers for the set of features we have chosen.)

$$\epsilon(f) = \underbrace{\epsilon(f) - \min_{f^* \in \mathcal{F}} \epsilon(f^*)}_{\text{estimation error}} + \underbrace{\min_{f^* \in \mathcal{F}} \epsilon(f^*)}_{\text{approximation error}}$$

We could maybe correct estimation error by getting more training data. More generally, we often refer to estimation error as **variance**.

We could maybe correct approximation error by choosing a better $\mathcal{F}$.

Bias-Variance Tradeoff

Let $\mathcal{F}$ denote the set of all possible classifiers under consideration. (E.g., all linear classifiers for the set of features we have chosen.)

$$\epsilon(f) = \underbrace{\epsilon(f) - \min_{f^* \in \mathcal{F}} \epsilon(f^*)}_{\text{estimation error}} + \underbrace{\min_{f^* \in \mathcal{F}} \epsilon(f^*)}_{\text{approximation error}}$$

We could maybe correct estimation error by getting more training data. More generally, we often refer to estimation error as **variance**.

We could maybe correct approximation error by choosing a better $\mathcal{F}$. More generally, we often refer to approximation error as **bias**.

References I

Hal Daume. *A Course in Machine Learning (v0.9)*. Self-published at http://ciml.info/, 2017.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.