Experimentation

October 27, 2015
Generalization

• We want to know how a predictor $h$ will perform in general.
• What do you mean in general?
  – “Average” behavior for all possible inputs (e.g., sentences, DNA sequences, corpora, ...), even the ones we don’t have in our training/test data

$$\mathbb{E}_{p(x,y)} \text{cost}(h(x), y)$$
Experimentation

• That expectation can’t be computed
  – Rather than looking at all possible inputs (maybe infinite! Maybe huge!), look at a representative sample of inputs
  – Make inferences from these experiments about the rest of the population
  – Rough idea: if we do well on a representative sample, we will do well on the whole population

• Mathematics can show provide conditions under which these inferences will be true with high probability
Standard Methodology

• We want to compare at two predictors $h$ and $h'$ that differ in a well-defined way
  – Data used to train them
  – Algorithm used to train them
  – Training objective (e.g., conditional vs. joint)
  – Feature set used
  – Inference method (e.g., exact vs. approximate)
  – Decoding objective (e.g., MAP vs. MBR)
Which predictor is better?

That is, we would like to know whether:

$$\mathbb{E}_{p(x,y)}[\text{cost} \ (h(x), y)] < \mathbb{E}_{p(x,y)}[\text{cost} \ (h'(x), y)]$$

Unfortunately, we cannot generally know this! 😞
Which predictor is better?

That is, we would like to know whether:

\[
\mathbb{E}_{p(x,y)}[\text{cost} \ (h(x), \ y)] < \mathbb{E}_{p(x,y)}[\text{cost} \ (h'(x), \ y)]
\]

Unfortunately, we cannot generally know this! 😞

But we can know the following: 😊

**Test set:** \( \mathcal{T} = \{ x_i^*, y_i^* \}_{i=1}^{N^*} \)

\[
\frac{1}{N^*} \sum_{i=1}^{N^*} \text{cost} \ (h(x_i^*), y_i^*) < \frac{1}{N^*} \sum_{i=1}^{N^*} \text{cost} \ (h'(x_i^*), y_i^*)
\]
Other Scenarios

• We may want to compare more than two predictors
• We may want to compare more than one cost function
• We may be working with cost functions that are defined at the corpus level
  – F-measure, precision, recall, BLEU, ROUGE, etc.
Held-Out Test Sets

• **Number one rule:** Keep your training data out of your test data
• If this sounds simple, it is anything but
  – Selecting hyperparameters by looking at the test set scores
  – Every year *many* (most?) papers are published that violate this!
• **Standard recipe**
  – *Training data* (possibly further subdivided into training & tuning)
  – Held-out *development data* [use while developing system]
  – Blind *test data* [for publication only]
Held-Out Test Sets

• Years of experimentation with “blind” test sets means they aren’t “blind” any longer!
• Strategies for dealing with this
  – Periodic creation of new test community sets
  – Fix all parameters of development data, report on held-out test data [publication bias]
  – Cross-validation

• I’ll say it again: Using held-out test data is the single most important thing you can do to ensure your experiments give generalization insight
Generalization: Cross Validation

- Sample train/dev/test data from D
- k-fold cross validation
  - Select k train/dev/test splits
- In the limit: k=N, “leave-one-out” CV
  - If you have N training instances, run N experiments training on N-1 instances
- Pros
  - More statistical power
  - Better use of limited data resources
- Cons
  - Computationally expensive
  - Not terribly common in structured prediction
Oracles and Upper Bounds

• What is the best possible performance knowing something about the test set?
  – Up to, and including, the test set!

• Examples
  – Tuning hyperparameters or parameters on the test set
  – Using gold standard parse trees or NER labels for a downstream information extraction task

• Answers a different question than generalization: does my model have adequate “capacity”? 
Back to Generalization

- Is held-out data enough?
- How many samples do we need to make reliable inferences?
  - If you need to detect big differences, you need fewer samples
  - If you need to detect small differences, you need big samples
  - If you do lots of similar experiments looking for an effect, you’re more likely to hit one “by chance”- can we control for this (false discovery)

- This brings us to...
Statistical Hypothesis Testing

- **Statistical predictors != statistical evaluation**
  - You can do statistical evaluation of non-statistical predictors!
- **Hypothesis testing in one sentence:** How likely is that the behavior we’re seeing is due to chance?
- **Hypothesis testing is not magical**
  - $p$-values are not the probability your claim is wrong
  - At best, you find out what is the probability some pattern of results is due to chance
    - If the your results unlikely due to chance, this does not mean the hypothesis you formulated was true; converse is also true
Statistical Hypothesis Testing

• Formulate a null hypothesis $H_0$
  – Skeptical perspective: e.g., two experimental scenarios are the same
• Set a threshold with which we reject the null hypothesis, usually
  $\alpha \in \{0.05, 0.01, 0.001\}$
• What is the probability of the experimental observations, assuming the null hypothesis?
  $p < \alpha$  $H_0$
  – If $p < \alpha$, then we can reject $H_0$
Parameters & Statistics

\[ u_i \sim U_i, \quad i = [1, N] \]
\[ v_i = v(u_i), \quad (\text{ie.}, \quad v_i \sim V) \]

The mean (a parameter) is not a random variable; it is a real number.

\[ \mu_V = \mathbb{E}_{p(u)}[v(u)] = \int v(u) \cdot p(u) du \]

The sample mean (a statistic) is a function of \( u \), and therefore is a random variable.

\[ \hat{\mu}_V = \frac{1}{N} \sum_{i=1}^{N} v_i \]
Sampling Distribution

• A statistic, e.g. our sample mean

\[ \hat{\mu}_V = \frac{1}{N} \sum_{i=1}^{N} v_i \]

is a random variable.

• What distribution is it drawn from, i.e. can we say something about the following?

\[ \hat{\mu}_V \sim \text{Distribution}(\theta) \]
Sampling Distribution

• Under some weak assumptions, a central limit theorem tells us
  \[ \hat{\mu}_V \sim \mathcal{N} \left( \mu_V, \frac{\sigma^2_V}{N} \right) \]

• This is an awesome result! As \( N \) gets bigger, the expected deviation from the parameter of interest drops.
Standard Error

• What is the standard deviation of the sample mean?

\[ \sigma_{\hat{\mu}_V} \] parameter of sampling distribution

\[ \sigma_{\mu_V} \] parameter of global population

\[ \sigma_{\hat{\mu}_V} = \frac{\sigma_V}{\sqrt{N}} \]
Standard Error

- What is the standard deviation of the sample mean?

\[ \sigma_V \text{ parameter of global population} \]

\[ \sigma_{\hat{\mu}_V} \text{ parameter of sampling distribution} \]

\[ \sigma_{\hat{\mu}_V} = \frac{\sigma_V}{\sqrt{N}} \]

\[ \hat{\sigma}_V \text{ statistic: the sample standard deviation} \]

\[ \hat{\sigma}_V = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (u_i - \hat{\mu}_i)^2} \]
Standard Error

- We can now state the standard error

\[ \hat{\sigma}_{\mu_V} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (u_i - \hat{\mu}_i)^2} \frac{1}{\sqrt{N}} \]

- This idea of replacing the true distribution (which we cannot know) with samples is the same thing we did with Monte Carlo techniques.
Standard Deviations
Other Parameters/Statistics

• Any **generalized mean:**
  – min, median, ..., max

• Proportions
  – proportion of a population for which property P holds

• Other functions
  – BLEU score, F-measure, word error rate...

• Except for proportions, these statistics don’t have a closed form of the standard error
Bootstrap (Efron, 1979)

- Monte Carlo technique to estimate standard error of some statistic
- We have a sample of $N$ draws from $U$

$$\mathbf{u} = (u_1, u_2, \ldots, u_N)$$

- For $i=1$ to $B$, resample $N$ times from the empirical distribution of $\mathbf{u}$

$$\mathbf{u}^{(i)} = (u^{(i)}_1, u^{(i)}_2, \ldots, u^{(i)}_N)$$
• From the sequence of bootstrap samples estimate the standard error

\[
\hat{\sigma}_\theta^{(boot)} = \sqrt{\frac{1}{B - 1} \sum_{i=1}^{B} \left( \hat{\theta}_{V,u(i)} - \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_{V,u(i)} \right)^2}
\]

\[
= \sqrt{\sum_{i=1}^{B} \left( \hat{\theta}_{V,u(i)} - \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_{V,u(i)} \right)^2}
\]

\[
\sqrt{B - 1}
\]

\[
\sigma_\theta \approx \hat{\sigma}_\theta \approx \hat{\sigma}_\theta^{(boot)}
\]

(When \( \theta_V = \mu_V \),

\[
\hat{\sigma}_V = \sigma_V / \sqrt{N}
\]