Announcements

- New date for exam: May 20
- HW4 typo: See Kevin's email

Outline for today

- Last comments on LDA
- How do you measure performance?
  The role of data
  Hard vs. soft decisions
  Trusting your performance estimates
  Comparing classifiers (trust)
Finishing up LDA ....

Connecting LDA classifier design & feature extraction

For linear classifiers with Gaussians:

\[ p(x | w_i) \sim N(\mu_i, \Sigma) \]

same for all

\[ w_i = \text{argmax} \ x^T \Sigma^{-1} \mu_i + k_i \]

for two classes

\[ x^T \Sigma^{-1} (\mu_1 - \mu_2) \geq t \]

linear transformation

of x project to 1 dimension

LoA result:

\[ Bu = d \Sigma \mathbf{u} \}

if \( W \) is invertible

\[ W^T Bu = d \mathbf{u} \]
Measuring Performance

1. The Role of Data

Assume "test" means independent (new) data

Important: The test sample is random

Why evaluate?
- To sell your stuff (commercially or academically)
- To find the right model (complexity, feature set, transformation, ...)

| Dev Design | Test | Data |
Idea of Independent Data:
- non-overlapping in terms of samples
- non-overlapping in terms of characteristics that are not controlled in test

Idea of Representative Data:
- contains different variations of things expected in test

Example 1: speaker-independent word recognition
  indep: different speakers in train/test
  rep: multiple speakers in each set

Example 2: whale identification
  indep: record in different locations
  rep: whale personality, time of year, age, gender, species, climate, non-whale sounds
2. Hard vs. Soft Decisions

Hard decisions: examples
\[ \hat{y} = \arg \max_w \ p(w|x) \] classification

\[ \hat{y} = \frac{1}{C} \sum_{i=1}^{C} x_i \] regression

Soft decision: examples
\[ p(w|x) \] for user to decide threshold
\[ p(\hat{y}|x) \] as a feature

Different Evaluation Criteria

- Error Averaging
- Risk Averaging (uses unlabeled data)
- ROC/DET Curve
- Cross-Validation
- Entropy
Error Averaging

\[ \hat{E} = \frac{1}{n} \sum_{i=1}^{n} C(y_i, \hat{y}_i) + \text{cost of deciding } \hat{y}_i \text{ when } y_i \text{ true} \]

Examples:
- Regression: \( C(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2 \)
- Classification: \( C(y_i, \hat{y}_i) = \begin{cases} 1 & y_i \neq \hat{y}_i \\ 0 & \text{otherwise} \end{cases} \)

For \( x_i \) i.i.d. d train/test independent
\( \hat{E} \) is typically unbiased and consistent

ROC/DET/P-R curves - Binary problems

Step 1: error counting for a particular threshold

<table>
<thead>
<tr>
<th>( \hat{y} )</th>
<th>true</th>
<th>false</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>( \text{N}_00 )</td>
<td>( \text{N}_01 )</td>
</tr>
<tr>
<td>1</td>
<td>( \text{N}_10 )</td>
<td>( \text{N}_11 )</td>
</tr>
</tbody>
</table>

\( n_0 \) total negatives, \( n_1 \) total positives

\( \text{DET curve:} \)

\[ P_m = \frac{n_{01}}{n_1}, \quad P_s = \frac{n_{10}}{n_0} \]

Precision - Recall

\[ P = \frac{n_{11}}{n_1}, \quad R = \frac{n_{11}}{n_0} \]
Cross Entropy:

\[ H_e(w|X) = -\frac{1}{n} \sum_{x} \sum_{\omega} p_e(\omega;x) \log p_m(\omega;w|x) \]

- empirical distribution of test data
- model
- test data
- \( g(x;w) \)
- \( \# \text{test samples} \)

Intuition: Oracle \( p_m(\omega;w|x_i) = 1 \) for \( w = w_i \)
- decent classifier \( p_m(\omega;w|x_i) \approx 0.8 \) most of the time
- lousy classifier \( p_m(\omega;w|x_i) = 0.1 \)

Normalized cross entropy:

\[ \frac{H(w) - H(w|x)}{H(w)} \]

range (ideally) is \( [0, 1] \)

- bad \( \Rightarrow \) poor
- good
3. Trusting your performance estimate

Which classifier would you buy?

A: $\hat{p}_A = 0$ (testing on 10 samples)

B: $\hat{p}_B = 0.01$ (testing on 1000 samples)

Check the confidence intervals

With 95% confidence, performance will be in the region:

A: $(0, 0.3)$

B: $(0, 0.015)$

*There is a 5% chance that the true performance is outside this range.

How do we find $(E_{low}, E_{high})$? $x^2$ comes

$P(E_{low} < \hat{E} < E_{high}) \geq x\%$

For error counting, $\hat{E}$ can be described as a binomial
Use the figure below (DHS Figure 9.10) to answer the following questions about confidence intervals.

1. If you have 100 samples and you observed 5 errors, what is the 95% confidence interval for the error rate?

\[ \hat{p} = 0.05 \quad (0.02, 0.1) \]

2. You anticipate a 90% classifier accuracy rate. How many samples would you need to have to be 95% confident that accuracy is greater than 85% correct if your estimate was 90%?

either \( 0.9 \) acc or \( 0.1 \) error OK

250 samples — biggest on graph w/o going \( > 0.15 \)

\( < 0.85 \)

**FIGURE 9.10.** The 95% confidence intervals for a given estimated error probability \( \hat{p} \) can be derived from a binomial distribution of Eq. 38. For each value of \( \hat{p} \), the true probability has a 95% chance of lying between the curves marked by the number of test samples \( n' \). The larger the number of test samples, the more precise the estimate of the true probability and hence the smaller the 95% confidence interval.
Distributions for finding confidence intervals:

- **Binomial**: good for small n, k
  \[ \hat{p} = k/n \]

For large n, can approximate with Poisson (esp. for small k) else with Gaussian

Detect 1 error in 100 trials

90% one-sided confidence interval

- **Binomial** \((0, 0.039)\) \(k = 1\) \(n = 100\)
- **Poisson** \((0, 0.038)\)
- **Gaussian** \((0, 0.023)\)