CS 440/ECE 448 Lecture 11: Exam 1 Review

Outline

- Exam administration
- Exam topics
- Solution methods for posted practice exam

Exam administration

- Where & when: Lincoln Hall Theater, 1:00pm, Mon Feb 16
- How: Paper exam w/ attached formula sheet and scratch paper
- What to bring:
  - pencils & erasers or pens
  - One 8.5x11 sheet of handwritten notes, front & back
  - ID
- Don’t bring:
  - Calculators or computers
  - Textbook or printed notes

Grading

- Each TA will grade one problem part
- Partial credit will be possible, using a rubric designed by the TA
- Regrades will be possible only if points can be changed while remaining fair to all other students in the class

Recommended approach to studying

- Do every variant of every daily quiz at least once (regenerating a variant will give it new numerical values, which can be helpful for practice)
- Make sure you understand the equations behind MP1, MP2, and MP3. The exam will not cover coding, but it will cover those equations.
- Then take the practice exam.
  - Treat the quizzes and MPs as your training set
  - Treat the practice exam as your dev set
  - Then you should have a pretty good idea how well you'll do on the test set

Outline

- Exam administration
- Exam topics
- Solution methods for posted practice exam

Topics included on the exam

- Technically: Everything we've covered so far
- In practice: Not including lectures 1 (intro), 2 (probability = background for the rest), 7 (fairness = interesting special case of decision theory, but you don't need to know the jargon)

Topics included on the exam

- Decision theory
  - MPE=MAP, confusion matrix, precision, recall, specificity, sensitivity
  - Naïve Bayes, unigrams, bigrams, ML parameters, Laplace smoothing
- Bayes networks
  - Space complexity, inference, independence, conditional independence
  - HMMs, Viterbi algorithm
- Learning
  - Risk, empirical risk, overtraining, train, dev, test
  - Linear regression, MSE, gradient descent, stochastic gradient descent

Decision Theory

Bayes Error Rate

$$= \sum_{x} P(x)\left(1 - \max_{y} P(y|x)\right)$$

$$\text{Precision} = P(Y = 1|f(X) = 1)$$
$$\text{Recall} = P(f(X) = 1|Y = 1)$$

$$\text{Specificity} = P(f(X) = 0|Y = 0)$$
$$\text{Sensitivity} = P(f(X) = 1|Y = 1)$$

Confusion Matrix

Correct Label: \ Classified As:	0	1
0	TN	FP
1	FN	TP

Naïve Bayes

- Naïve Bayes classifier:

$$f(x) = \text{argmax}(\log P(Y = y) + \log P(X = x|Y = y))$$

$$\log P(X = x|Y = y) \approx \sum_{i=1}^{n} \log P(W = w_i|Y = y)$$

- Laplace Smoothing w/OOV:

$$P(W = w|Y = y) = \begin{cases} \dfrac{k + \text{Count}(w, y)}{N + k(M + 1)} & \text{in} - \text{vocabulary} \\ \dfrac{k}{N + k(M + 1)} & \text{OOV} \end{cases}$$

- Laplace smoothing w/o OOV:

$$P(W = w|Y = y) = \frac{k + \text{Count}(w, y)}{N + kM}$$

Bayes networks

- Bayesian network: A better way to represent knowledge
  - Reduces space complexity from $\mathcal{O}\{v^n\}$ to $\mathcal{O}\{nv^p\}$ -- huge if $n \gg p$
  - Does not reduce time complexity without extra assumptions
- Key ideas: Independence and Conditional independence
  1. Shared ancestor $\Longrightarrow$ dependent unless ancestor known
  2. Shared descendant $\Longrightarrow$ dependent unless descendant unknown
  3. No shared ancestor or descendent $\Longrightarrow$ independent

HMMs

- Review: Bayesian classifier, Bayesian networks

$$f(x) = \underset{y}{\operatorname{argmax}}\, P(Y = y | X = x)$$

- HMM: Probabilistic reasoning over time

$$\pi_i = P(Y_1 = i)$$
$$a_{i,j} = P(Y_t = j | Y_{t-1} = i)$$
$$b_j(\boldsymbol{x}_t) = P(X_t = \boldsymbol{x}_t | Y_t = j)$$

- Viterbi algorithm

$$\textbf{SCORE:}\ v_t(j) = \max_i v_{t-1}(i) + \log a_{i,j} + \log b_j(\boldsymbol{x}_t)$$

$$\textbf{BACKPOINTER:}\ \psi_t(j) = \underset{i}{\operatorname{argmax}}\, v_{t-1}(i) + \log a_{i,j} + \log b_j(\boldsymbol{x}_t)$$

Learning

- **Learning:** Given $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, find the function $f(X)$ that minimizes some measure of risk.
- **Empirical risk**, a.k.a. training corpus error:

$$\mathcal{R}_{\text{emp}} = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$$

- **True risk,** a.k.a. expected test corpus error:

$$\mathcal{R} = \text{E}[\ell(Y, f(X))] = \mathcal{R}_{\text{emp}} + \mathcal{R}_{\text{generalization}}$$

- **Early Stopping**: Stop when error rate on the dev set reaches a minimum, then test on test set to get an unbiased estimate of true risk

Linear regression

- Definition of linear regression

$$f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{w}$$

- Mean-squared error

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i\,, \qquad \mathcal{L}_i = \frac{1}{2}\epsilon_i^2, \qquad \epsilon_i = f(\boldsymbol{x}_i) - y_i$$

- Gradient descent

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \frac{\partial \mathcal{L}}{\partial \boldsymbol{w}}, \qquad \frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \boldsymbol{x}_i$$

- Stochastic gradient descent

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \frac{\partial \mathcal{L}_i}{\partial \boldsymbol{w}}, \qquad \frac{\partial \mathcal{L}_i}{\partial \boldsymbol{w}} = \epsilon_i \boldsymbol{x}_i$$

Outline

- Exam administration
- Exam topics
- Solution methods for posted practice exam

1(a): the pastry alarm clock

(13 points) A Bayes classifier orders a pastry, then computes the time of day that has the highest *a posteriori* probability given the type of pastry. In terms of the parameters P_{AM}, P_{PM}, $P(F|AM)$, $P(F|PM)$, $P(M|AM)$, and/or $P(M|PM)$, what is the error rate of such a classifier? Your answer may contain explicit summations, minimizations, and/or maximizations; please specify the variable(s) over which you are summing, maximizing, and/or minimizing.

Solution method

$$\text{Bayes Error Rate} = \sum_x P(x)\left(1 - \max_y P(y|x)\right)$$

$$= \sum_x \left(P(x) - \max_y P(x)P(y|x)\right)$$

$$= \sum_x \sum_{y \neq argmax P(y)P(x|y)} P(y)P(x|y)$$

$$P(\text{Error}) = \sum_{P \in \{F,M\}} \min_{T \in \{AM,PM\}} P_T P(P|T)$$

$$= \sum_{P \in \{F,M\}} \sum_{T \neq \arg\max P_T P(P|T)} P_T P(P|T)$$

1(b) Laplace smoothing

(13 points) In the past 30 days you've awakened before noon 25 times, after noon 5 times. You received a fruit pastry before noon 16 times & after noon once; you received a meat pastry before noon 9 times and after noon 4 times. Assuming that all pastries have either meat or fruit but not both (any other option is impossible), and in terms of a Laplace smoothing parameter k, estimate P_{AM}, P_{PM}, $P(F|AM)$, $P(M|AM)$, $P(F|AM)$, and $P(F|PM)$.

Solution method

Laplace smoothing w/o OOV:

$$P(W = w|Y = y) = \frac{k + \text{Count}(w, y)}{N + kM}$$

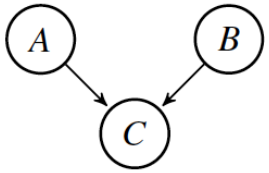
$$P_{AM} = \frac{25+k}{30+2k}, \quad P_{PM} = \frac{5+k}{30+2k}$$

$$P(F|AM) = \frac{16+k}{25+2k}, \quad P(M|AM) = \frac{9+k}{25+2k}$$

$$P(F|PM) = \frac{1+k}{5+2k}, \quad P(M|PM) = \frac{4+k}{5+2k}$$

2(a): Independence & Conditional Indep

Consider three binary events A, B, and C, related by the Bayes network shown here:



The parameters of this Bayes network are given by the unknown constants u through z, as follows:

$$P(A) = u, \quad P(B) = v$$
$$P(C|A,B) = w, \quad P(C|A,\neg B) = x$$
$$P(C|\neg A,B) = y, \quad P(C|\neg A,\neg B) = z$$

(a) (10 points) Are events A and B independent, conditionally independent given knowledge of C, both, or neither? Explain your answer.

2(a): Solution

- No ***unknown ancestors***, so they are independent
- If C is known, then they have a ***known descendent***, so they are not conditionally independent given C

2(b): Probability & Conditional Probability

$$P(A) = u, \quad P(B) = v$$
$$P(C|A,B) = w, \quad P(C|A,\neg B) = x$$
$$P(C|\neg A,B) = y, \quad P(C|\neg A,\neg B) = z$$

(10 points) Suppose C is unknown. Find the four probabilities $P(A,B)$, $P(A,\neg B)$, $P(\neg A,B)$, and $P(\neg A,\neg B)$.

2(b) Solution method

$$P(X,Y) = P(X|Y)P(Y)$$

$$\begin{aligned} P(A) = u, \quad & P(B) = v \\ P(C|A,B) = w, \quad & P(C|A,\neg B) = x \\ P(C|\neg A,B) = y, \quad & P(C|\neg A,\neg B) = z \end{aligned}$$

$$\begin{aligned} P(A,B) &= uv \\ P(A,\neg B) &= u(1-v) \\ P(\neg A,B) &= (1-u)v \\ P(\neg A,\neg B) &= (1-u)(1-v) \end{aligned}$$

2 c: Conditional Probability

$$P(A) = u, \quad P(B) = v$$
$$P(C|A,B) = w, \quad P(C|A,\neg B) = x$$
$$P(C|\neg A,B) = y, \quad P(C|\neg A,\neg B) = z$$

(c) (10 points) Suppose C is known to be true. Find $P(C)$, and in terms of $P(C)$, find $P(A,B|C)$, $P(A,\neg B|C)$, $P(\neg A,B|C)$, and $P(\neg A,\neg B|C)$.

Solution:

$$P(C) = uvw + u(1-v)x + (1-u)vy + (1-u)(1-v)z$$
$$P(A,B|C) = \frac{uvw}{P(C)}$$
$$P(A,\neg B|C) = \frac{u(1-v)x}{P(C)}$$
$$P(\neg A,B|C) = \frac{(1-u)vy}{P(C)}$$
$$P(\neg A,\neg B|C) = \frac{(1-u)(1-v)z}{P(C)}$$

3: Viterbi algorithm

You've been hired as a night watchman at a nuclear power plant. Let $Y_t = 1$ if the power plant is overheating at time t, otherwise $Y_t = 0$, and suppose that $Y_0 = 0$. Let $X_t = 1$ if the reactor warning light is blinking at time t, otherwise $X_t = 0$. Define $a_{ij} = P(Y_{t+1} = j | Y_t = i)$, and $b_{ij} = P(X_t = j | Y_y = i)$. Suppose there is some normalizer, z, and base, c, such that $\log_c(a_{ij}/z) = \log_c(b_{ij}/z) = 0$ for $i = j$, but for $i \neq j$, $\log_c(a_{ij}/z) = -2$ and $\log_c(b_{ij}/z) = -3$. Suppose that $\{X_1, \ldots, X_9\} = \{1,0,0,1,1,1,0,0,1\}$. Draw a trellis specifying the numerical values of $v_t(i) = \max \log_c P(Y_1, \ldots, Y_t = i | X_1, \ldots, X_t)/z$ for $i \in \{0,1\}$ and $1 \leq t \leq 9$, and specify a sequence of state variables $\{Y_1, \ldots, Y_9\}$ that maximizes the score. If there are more than one state sequences with the same maximum score, you only need to provide one of them.

3: Solution method

SCORE: $v_t(j) = \max_i v_{t-1}(i) + \log a_{i,j} + \log b_j(\boldsymbol{x}_t)$

BACKPOINTER: $\psi_t(j) = \operatorname{argmax}_i v_{t-1}(i) + \log a_{i,j} + \log b_j(\boldsymbol{x}_t)$

$$\log a_{ij}/z = \begin{cases} 0 & i = j \\ -2 & i \neq j \end{cases}, \qquad \log b_{ij}/z = \begin{cases} 0 & i = j \\ -3 & i \neq j \end{cases},$$

t	1	2	3	4	5	6	7	8	9
X_t	1	0	0	1	1	1	0	0	1
$v_t(0)$	$\leftarrow -3$	$\leftarrow -3$	$\leftarrow -3$	$\leftarrow -6$	$\leftarrow -9$	$\swarrow -10$	$\swarrow -7$	$\leftarrow -7$	$\leftarrow -10$
$v_t(1)$	$\nwarrow -2$	$\leftarrow -5$	-8	$\nwarrow -5$	$\leftarrow -5$	$\leftarrow -5$	$\leftarrow -8$	$\leftarrow -11$	$\nwarrow -9$

4(a) Gradient descent

Consider a linear regression model $f(\mathbf{x}_i) = \mathbf{w}^T \mathbf{x}_i$ where $\mathbf{x}_i$ is an m-dimensional vector drawn from a size-n training corpus, $1 \leq i \leq n$. Suppose you want to choose the vector $\mathbf{w}$ to minimize $\mathscr{L}_{\text{train}}$, defined as

$$\mathscr{L}_{\text{train}} = \frac{1}{n} \sum_{i=1}^{n} \left| \ln \left(\frac{f(\mathrm{x}_i)}{y_i} \right) \right|,$$

where ln is the natural logarithm, and y_i is a real-valued scalar for $1 \leq i \leq n$.

(a) (12 points) Find the gradient $\frac{\partial \mathscr{L}_{\text{train}}}{\partial \mathbf{w}}$.

4(a) Solution method: use the chain rule

$$\mathscr{L}_{\text{train}} = \frac{1}{n}\sum_{i=1}^{n}\left|\ln\left(\frac{f(\mathrm{x}_i)}{y_i}\right)\right|,$$

where ln is the natural logarithm, and y_i is a real-valued scalar for $1 \leq i \leq n$.

(a) (12 points) Find the gradient $\frac{\partial \mathscr{L}_{\text{train}}}{\partial \mathbf{w}}$.

Solution:

$$\frac{\partial \mathscr{L}_{\text{train}}}{\partial \mathbf{w}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i}{f(\mathbf{x}_i)}\operatorname{sign}\left(\ln\left(\frac{f(\mathrm{x}_i)}{y_i}\right)\right)$$

4(b) Learning

(12 points) Suppose you also have a development corpus, with tokens numbered $n+1 \leq i \leq n+j$, and a test corpus, with tokens numbered $n+j+1 \leq i \leq n+j+k$. The losses in these corpora are

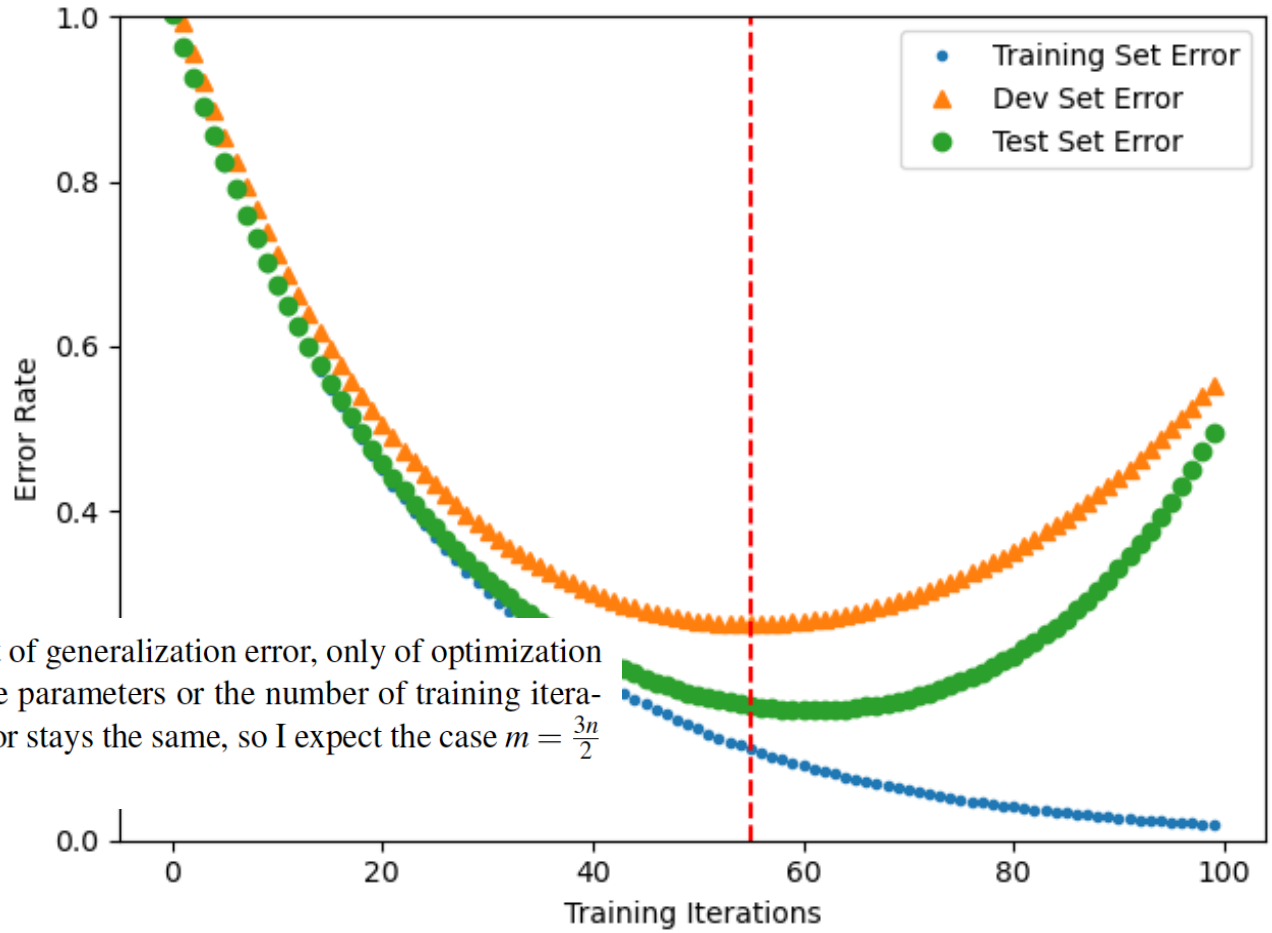
$$\mathscr{L}_{\text{dev}} = \frac{1}{j} \sum_{i=n+1}^{n+j} \left| \ln \left(\frac{f(\mathrm{x}_i)}{y_i} \right) \right|,$$

$$\mathscr{L}_{\text{test}} = \frac{1}{k} \sum_{i=n+j+1}^{n+j+k} \left| \ln \left(\frac{f(\mathrm{x}_i)}{y_i} \right) \right|.$$

Your client, a hardware manufacturer, is very clever about finding things to measure. They propose the following experiment: (1) increase m by finding something else to measure about each product, (2) re-train the new $\mathbf{w}$, finding the value that minimizes $\mathscr{L}_{\text{train}}$, (3) measure $\mathscr{L}_{\text{dev}}$ and $\mathscr{L}_{\text{test}}$, (4) repeat. Do you expect $\mathscr{L}_{\text{train}}$ to be smallest when $m = 1$, when $m = \frac{n}{2}$, or when $m = \frac{3n}{2}$? Why?

4(b) Solution method

$$\mathcal{R} = \mathrm{E}[\ell(Y, f(X))]$$
$$= \mathcal{R}_{\text{emp}} + \mathcal{R}_{\text{generalization}}$$



Solution: $\mathscr{L}_{\text{train}}$ does not include any component of generalization error, only of optimization error. When you increase the number of trainable parameters or the number of training iterations, optimization error always either decreases or stays the same, so I expect the case $m = \frac{3n}{2}$ will have the smallest $\mathscr{L}_{\text{train}}$.

Try the quiz!

- Today’s quiz is about word2vec
- It is not relevant to the exam!
- But it is is relevant to MP4
- Anyway, give it a try