Name: ____________________________________ NetID: _____________

S&DS 365 / 565

Introductory Machine Learning

Midterm Exam #2 (Practice)

Tuesday, November 19, 2019

Complete all of the problems. You are allowed one double-sided (8.5×11) sheet of paper with notes. No electronic devices, including calculators.

For your reference, recall the following distributions:

$$\text{Normal}(\mu, \sigma^2): \qquad p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}, \quad x \in \mathbb{R}$$

$$\text{Beta}(\alpha, \beta): \qquad p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \theta \in [0,1]$$

$$\text{Dirichlet}(\alpha_1, \ldots, \alpha_K): \qquad p(\theta) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_k^{\alpha_k - 1}, \quad \sum_{k=1}^K \theta_k = 1, \;\; \theta_k \geq 0$$

$$\text{Bernoulli}(\theta): \qquad \mathbb{P}(z) = \theta^z (1-\theta)^{(1-z)}, \quad z \in \{0, 1\}$$

$$\text{Binomial}(n, p): \qquad \mathbb{P}(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0, 1, \ldots, n\}$$

1	2	3	4	5	total
20	10	10	10	10	60

1. *True or False, Yes or No?* (20 points)

   Indicate whether each of the following statements is true or false, by circling your answer.

class-based

vs

embedding-based

In a class-based bigram language model, the probability of the next word is

$$p(w_n \,|\, w_{n-1}) = \overbrace{p(w_n \,|\, c(w_n))}\,\overbrace{p(c(w_n) \,|\, c(w_{n-1}))}$$

curr prev curr class curr class curr class prev

e.g bus e.g noun e.g noun e.g verb

where $c(w)$ is the class of word w.

In an embedding-based bigram language model, the probability of the next word is

$$p(w_n \,|\, w_{n-1}) = \frac{\exp(\phi(w_n)^T \phi(w_{n-1}))}{\sum_{v \in \text{Vocabulary}} \exp(\phi(v)^T \phi(w_{n-1}))}$$

where $\phi(w) \in \mathbb{R}^{100}$ is embedding vector for word w.

YES NO (a) The classes can be determined by bottom-up clustering.

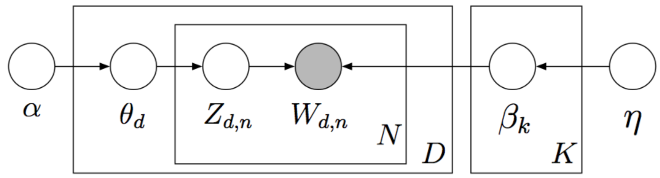
YES NO (b) As the number of classes decreases, the training likelihood decreases.

YES NO (c) Words that co-occur with each other will tend to be placed in the same class.

YES NO (d) The embedding vectors can be obtained by maximizing the training likelihood using stochastic gradient descent.

YES NO (e) Both of these are mixture models with latent variables.

The following questions concern latent Dirichlet allocation (LDA) topic model



where $\theta_d \sim \text{Dirichlet}(\alpha)$ are the per-document topic proportions, $Z_{d,n} \sim \text{Multi}(\theta_d)$ are the per-word topic assignments, $\beta_k \sim \text{Dirichlet}(\eta)$ are the topics, and $W_{d,n} \sim \text{Multi}(\beta_{Z_{d,n}})$ are the observed words.

YES NO (a) The objective of posterior inference is to estimate the conditional probability of θ_d, $Z_{d,n}$, and β_k given a corpus of documents.

YES NO (b) The words in a document are generated independently.

YES NO (c) Reordering the documents in the training corpus does not change the posterior distribution over the latent variables.

YES NO (d) Reordering the words in each document in the training corpus does not change the posterior distribution over the latent variables.

YES NO (e) Conditioned on all of the per-word topic assignments $Z_{d,n}$, the posterior distribution over θ_d is a single Dirichlet distribution.

The following questions concern the t-SNE method for visualizing word embeddings.

YES NO (a) The method is used to visualize high dimensional embeddings.

YES NO (b) The goal of the algorithm is to preserve all pairwise distances between words in the high dimensional embedding space.

YES NO (c) The algorithm attempts to place words nearby in the 2-dimensional visualization that are very close together in the high dimensional space.

YES NO (d) The method gives another representation of the data that is commonly used in other machine learning algorithms.

YES NO (e) The method is based on the use of latent variable models.

YES NO (f) The method uses Bayesian inference.

YES NO (g) The visualization is optimized using stochastic gradient descent.

2. *Bayesian inference* (10 points)

Suppose X is a Bernoulli(θ) random variable and we observe data $D_n = \{x_1, \ldots, x_n\}$. Suppose the prior distribution on θ is Beta(α, β). Let $s = \sum_{i=1}^{n} x_i$ be the number of "successes".

(a) What is the posterior distribution of θ given D_n?

(b) What is the posterior mean?

3. *Embeddings* (10 points)

(a) Define the pointwise mutual information

(b) Consider the analogy "puppy is to dog as ? is to cat" Use $\phi(\text{puppy})$, $\phi(\text{dog})$, and $\phi(\text{cat})$ to denote the embeddings of these three words. Write down the missing word as the solution to an optimizition problem.

4. *Neural nets* (10 points)

Consider a function $f(x) = \sigma(W_2\sigma(W_1x + b_1) + b_2)$ where σ is an activation function. Considering f as a neural network, suppose that the input dimension is $d = 2$, and there are two neurons in each hidden layer.

(a) Draw a picture of f as a network.

(b) Suppose the activation function is $\sigma(x) = \text{relu}(x)$ is the rectified linear unit, and that

$$W_1 = \begin{pmatrix} 2 & 1 \\ 2 & 2 \end{pmatrix} \quad b_1 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$$
$$W_2 = \begin{pmatrix} 1 & -1 \end{pmatrix} \quad b_2 = -1.$$

Compute the value of $f((2, -2)^T)$.

5. *Code* (10 points)

(a) What is the value of the following Python expression?

```
np.mean([i+1 for i in range(3)])
```

(b) Explain in a couple of sentences what the function `mystery` defined below does:

```
def mystery(X, Y):
  n, d = X.shape

  W1 = .1*np.random.randn(d,10)
  b1 = np.random.randn(1,10)
  W2 = .1*np.random.randn(10,1)
  b2 = np.random.randn(1,1)

  for i in range(1000):
      h = np.dot(X, W1) + b1
      Yhat = np.dot(h, W2) + b2
      dloss = Yhat - Y
      dloss /= num_examples
      dW2 = np.dot(h.T, dloss)
      db2 = np.sum(dloss)
      dh = np.dot(dloss, W2.T)
      dW1 = np.dot(X.T, dh)
      db1 = np.sum(dh)
      W1 += -.01 * dW1
      b1 += -.01 * db1
      W2 += -.01 * dW2
      b2 += -.01 * db2
  return Yhat
```