Introduction to Machine Learning, Spring 2016

Problem Set 6: Probabilistic models

Due: Monday, April 25, 2016 at 10pm (upload to NYU Classes.)

Important: *See problem set policy on the course web site.*

Instructions. You must show **all** of your work and be rigorous in your writeups to obtain full credit. Your answers to the below, plots, and all code that you write for this assignment should be uploaded to NYU Classes.

1. **Medical diagnosis**. You go for your annual checkup and have several lab tests performed. A week later your doctor calls you and says she has good and bad news. The bad news is that you tested positive for a marker of a serious disease, and that the test is 97% accurate (i.e. the probability of testing positive given that you have the disease is 0.97, as is the probability of testing negative given that you don't have the disease). The good news is that this is a rare disease, striking only 1 in 20,000 people. Why is it good news that the disease is rare? What are the chances that you actually have the disease?

2. **Naive Bayes**. In this problem you will show that naive Bayes corresponds to a linear classifier. Consider using a naive Bayes algorithm for binary prediction (two classes), where the features $x_1, \ldots, x_k$ are also binary valued. Let $\theta_c = \Pr(Y = c)$ and $\theta_{ci} = \Pr(X_i = 1 \mid Y = c)$ for $c \in \{0, 1\}$. It will be helpful to use the following form for the joint distribution:

$$\Pr(Y = 1, x_1, \ldots, x_k \,;\, \vec{\theta}) = \theta_1 \prod_{i=1}^{k} \theta_{1i}^{x_i} (1 - \theta_{1i})^{1 - x_i} \tag{1}$$

$$\Pr(Y = 0, x_1, \ldots, x_k \,;\, \vec{\theta}) = \theta_0 \prod_{i=1}^{k} \theta_{0i}^{x_i} (1 - \theta_{0i})^{1 - x_i} \tag{2}$$

   For a naive Bayes model given by parameters $\vec{\theta}$, demonstrate a weight vector $\mathbf{w}$ and offset b such that for any new example $\mathbf{x}$,

$$\arg\max_y \Pr(y \mid \mathbf{x} \,;\, \vec{\theta}) \quad = \quad \arg\max_y y \, (\mathbf{w} \cdot \mathbf{x} + b) \,,$$

   where $\vec{\theta}$ refers to all parameters, including both θ_c and θ_{ci}.

   Hint: Use Bayes' rule to obtain the posterior, and then take its logarithm (noticing that this is a monotonic transformation which does not change the argmax).

   Thus, if one had a sufficient amount of data, one would prefer to directly learn a linear model using logistic regression or a SVM rather than using naive Bayes, since the former consider a strictly larger hypothesis class than the latter. With limited numbers of training points (or settings where some features may be missing) naive Bayes may be preferable.

3. **Topic models.** In this question you will use an off-the-shelf implementation of LDA to get practice with learning topic models on real-world data, and to analyze various trade-offs that can be made during learning.

   (a) Prepare a corpus of documents from which you'll learn. You can find some already prepared text collections here:

   `http://archive.ics.uci.edu/ml/datasets/Bag+of+Words`

   However, we prefer that you be creative and construct your own!

   (b) Learn a latent Dirichlet allocation model on your corpus using default parameters. You can use any software package that you like. Two excellent options are:

- Mallet (`http://mallet.cs.umass.edu/`)
- Gensim (`http://radimrehurek.com/gensim/`)

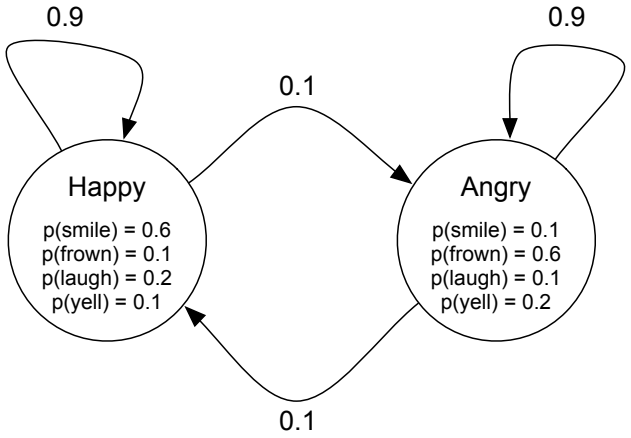
Qualitatively describe what topics are discovered.

(c) Re-run learning using varying numbers of topics (e.g., 5, 20, 100). Describe qualitatively the differences that you observe as the number of topics increases.

4. **Hidden Markov models**. Andy lives a simple life. Some days he is Angry and some days he is Happy. But he hides his emotional state, and so all we can observe is whether he smiles, frowns, laughs, or yells. Andy's best friend is utterly confused about whether Andy is actually happy or angry and decides to model his emotional state using a hidden Markov model.

   Let $X_d \in \{\text{Happy, Angry}\}$ denote Andy's emotional state on day d, and let $Y_d \in \{\text{smile, frown, laugh, yell}\}$ denote the observation made about Andy on day d. **Assume that on day 1 Andy is in the Happy state**, i.e. $X_1 = \text{Happy}$. Furthermore, assume that Andy transitions between states exactly once per day (staying in the same state is an option) according to the following distribution: $p(X_{d+1} = \text{Happy} \mid X_d = \text{Angry}) = 0.1$, $p(X_{d+1} = \text{Angry} \mid X_d = \text{Happy}) = 0.1$, $p(X_{d+1} = \text{Angry} \mid X_d = \text{Angry}) = 0.9$, and $p(X_{d+1} = \text{Happy} \mid X_d = \text{Happy}) = 0.9$.

   The observation distribution for Andy's Happy state is given by $p(Y_d = \text{smile} \mid X_d = \text{Happy}) = 0.6, p(Y_d = \text{frown} \mid X_d = \text{Happy}) = 0.1, p(Y_d = \text{laugh} \mid X_d = \text{Happy}) = 0.2$, and $p(Y_d = \text{yell} \mid X_d = \text{Happy}) = 0.1$. The observation distribution for Andy's Angry state is $p(Y_d = \text{smile} \mid X_d = \text{Angry}) = 0.1, p(Y_d = \text{frown} \mid X_d = \text{Angry}) = 0.6, p(Y_d = \text{laugh} \mid X_d = \text{Angry}) = 0.1$, and $p(Y_d = \text{yell} \mid X_d = \text{Angry}) = 0.2$. **All of this is summarized in the following figure**:



   Be sure to show all of your work for each of the questions below!

   (a) What is $p(X_2 = \text{Happy})$?

   (b) What is $p(Y_2 = \text{frown})$?

   (c) What is $p(X_2 = \text{Happy} \mid Y_2 = \text{frown})$?

   (d) What is $p(Y_{80} = \text{yell})$?

   (e) Assume that $Y_1 = Y_2 = Y_3 = Y_4 = Y_5 = \text{frown}$. What is the most likely sequence of the states? That is, compute the MAP assignment $\arg\max_{x_1,\ldots,x_5} p(X_1 = x_1, \ldots, X_5 = x_5 \mid Y_1 = Y_2 = Y_3 = Y_4 = Y_5 = \text{frown})$.