Introduction to Machine Learning, Fall 2013

Problem Set 5: Bayesian methods

Due: Thursday, November 21, 2013 by 11am (in class, *before* class begins)

Important: *See problem set policy on the course web site.* **You must show all of your work and be rigorous in your writeups to obtain full credit.**

---

1. (10 points) **Medical diagnosis**

   You go for your yearly 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 99% accurate (i.e. the probability of testing positive given that you have the disease is 0.99, 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 10,000 people. Why is it good news that the disease is rare? What are the chances that you actually have the disease?

2. (10 points) **Fitting a naive Bayes spam filter by hand**

   Consider a naive Bayes model (where the class takes two states) for spam classification with the vocabulary $V =$ "secret", "offer", "low", "price", "valued", "customer", "today", "dollar", "million", "sports", "is", "for", "play", "healthy", and "pizza". We have the following example spam messages: (b1) "million dollar offer", (b2) "secret offer today", (b3) "secret is secret". We also have the following normal messages: (g1) "low price for valued customer", (g2) "play secret sports today", (g3) "sports is healthy", (g4) "low price pizza". Give the maximum likelihood estimates (MLE) for the following parameters: $\theta_{\text{spam}}, \theta_{\text{secret}|\text{spam}}, \theta_{\text{secret}|\text{non-spam}}, \theta_{\text{sports}|\text{non-spam}}$, and $\theta_{\text{dollar}|\text{spam}}$.

3. (10 points) **Missing features in naive Bayes**

   Consider a naive Bayes model given by:

$$\begin{aligned} \Pr(Y = y, X_1 = x_1, \ldots, X_n = x_n; \vec{\theta}) &= \Pr(Y = y; \vec{\theta}) \prod_{i=1}^{N} \Pr(X_i = x_i \mid Y = y; \vec{\theta}) \quad (1) \\ &= \theta_y \prod_{i=1}^{N} \theta_{x_i|y}, \quad (2) \end{aligned}$$

   where $\vec{\theta}$ refers to the parameters of the model. Let $\mathcal{O} \subseteq \{1, \ldots, N\}$ denote variables that are observed in a new instance. Show how to compute the posterior $\Pr(Y = y \mid \mathbf{X}_{\mathcal{O}} = \mathbf{x}_{\mathcal{O}})$ in running time $O(|\mathcal{O}|)$ (as opposed to $O(N)$), where $\mathbf{x}_O$ denotes the assignment to the variables that were observed (all other variables' values are unobserved).

4. (20 points) **LDA and naive Bayes are linear classifiers**

   (a) Suppose we have training data given by $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$. Consider the following learning algorithm for binary classification with features $\mathbf{X} \in \mathbb{R}^k$ and labels $Y \in \{0, 1\}$. First, let $\Pr(Y = 1) = \frac{1}{N} \sum_{i=1}^{N} y_i$ be the fraction of training examples

labeled 1. Then we find the maximum likelihood fit of two multi-variate Gaussian distributions to the data in each class, where the Gaussian distribution is given by

$$\Pr(\mathbf{x}; \mu_Y, \mathbf{\Sigma}_Y) = \frac{1}{(2\pi)^{k/2}|\Sigma_Y|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_\mathbf{Y})^T \Sigma_Y^{-1}(\mathbf{x} - \mu_\mathbf{Y})\right).$$

For example, the MLE estimates for the means are given by (for $Y \in \{0, 1\}$):

$$\hat{\mu}_Y = \frac{1}{|\{(x_i, y_i) \in \mathcal{D} : y_i = Y\}|} \sum_{(x_i, y_i) \in \mathcal{D}: y_i = Y} x_i.$$

This algorithm is known as linear discriminant analysis (LDA).[1]

In this problem, show that when the co-variance matrices are equal, i.e. $\Sigma_0 = \Sigma_1$, then maximum a posteriori (MAP) classification using this model is given by a linear discriminant function (**note:** we are *not* assuming that the co-variance matrices are diagonal, i.e. the X_i variables may not be conditionally independent given Y). Specifically, for any μ_0, μ_1 and Σ, demonstrate a weight vector $\mathbf{w}$ and offset b such that for any new example $\mathbf{x}$,

$$\arg\max_y \Pr(y \mid \mathbf{x}\,;\, \mu_0, \mu_1, \Sigma) \quad = \quad \arg\max_y y\,(\mathbf{w} \cdot \mathbf{x} + b)\,.$$

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).

(b) Show that the same holds for naive Bayes. In particular, 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:

$$\Pr(Y = 1, x_1, \ldots, x_k\,;\, \vec{\theta}) = \theta_1 \prod_{i=1}^{k} \theta_{1i}^{x_i}(1 - \theta_{1i})^{1-x_i} \tag{3}$$

$$\Pr(Y = 0, x_1, \ldots, x_k\,;\, \vec{\theta}) = \theta_0 \prod_{i=1}^{k} \theta_{0i}^{x_i}(1 - \theta_{0i})^{1-x_i} \tag{4}$$

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}.

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 LDA or 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) LDA and naive Bayes may be preferable.

[1] Note, this acronym is also used for Latent Dirichlet Allocation, which is a type of "topic model" and has nothing to do with linear discriminant analysis.