STAT 339: HOMEWORK 3 (BAYESIAN CLASSIFICATION AND REGRESSION)

UPDATED: DUE ON BLACKBOARD BY SUNDAY 3/26)

Instructions. Turn in your writeup and code to Blackboard as an archive file (e.g., `.zip`, `.tar`, `.gz`) by the start of class on Friday 3/3.

As always, you may use any language you like for the progamming components of this assignment — the tasks are stated in a language-neutral way. You may also use any typesetting software to prepare your writeup, but the final document should be a PDF. LaTeXis encouraged; a reproducible research format in which code is embedded into the document (e.g., knitr, RMarkdown, Jupyter or IPython Notebook) is even more encouraged.

All data is available at `http://colindawson.net/data/<name>.csv`.

1. Naive Bayes With Categorical Features

1. **Warmup I: Spam Filtering** (Adapted from BRML 10.5) This question concerns spam filtering. Each email is represented by a vector

$$\mathbf{x} = (x_1, ..., x_D)$$

where $x_d \in \{0, 1\}$. Each entry of the vector indicates if a particular symbol or word (out of D symbols/words in the vocabulary) appears in the email. The symbols/words are things like

$$money, cash, !!!, viagra, \ldots, etc.$$

so that, for example, $x_2 = 1$ if the word 'cash' appears in the email. The training dataset consists of a set of vectors along with the class label c, where $c = 1$ indicates the email is spam, and $c = 0$ indicates that it is not spam. Hence, the

Date: Last Revised: March 23, 2017.

training set consists of a set of pairs $(\mathbf{x}_n, c_n), n = 1, \ldots, N$. The naive Bayes model for the joint probability of the category and contents of email n is

$$p(c_n, \mathbf{x}_n) = p(c_n) \prod_{d=1}^{D} p(x_{nd} \mid c_n)$$

(a) Derive expressions for the parameters of this model in terms of the training data using maximum likelihood. Assume that the data is independent and identically distributed; that is that

$$p(c_1, ..., c_N, \mathbf{x}_1, ..., \mathbf{x}_N) = \prod_{n=1}^{N} p(c_n, \mathbf{x}_n)$$

Explicitly, the parameters are $(\pi, \theta_{01}, \ldots, \theta_{0D}, \theta_{11}, \ldots, \theta_{1D})$, where

$$\pi := p(c = 1)$$
$$\theta_{1d} := p(x_d = 1 \mid c = 1)$$
$$\theta_{0d} := p(x_d = 1 \mid c = 0)$$

(The same parameters are assumed to apply for every email of the same type, which is why n does not appear in their definitions.)

(b) Given a trained model (i.e., given estimated values of the π and θ parameters, and thus an estimate of the joint PMF $p(\mathbf{X}, \mathbf{c})$ (where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ and $\mathbf{c} = (c_1, \ldots, c_N)$), explain how to form a classifier that returns $p(c_{new} \mid \mathbf{x}_{new})$.

(c) If 'viagra' never appears in the spam training data, discuss what effect this will have on the classification for a new email that contains the word 'viagra'. Explain how you might counter this effect. Explain how a spammer might try to fool a naive Bayes spam filter.

2. **Warmup II: Customer Categorization** (Adapted from BRML 10.1) A local supermarket specializing in breakfast cereals decides to analyze the buying patterns of its customers. They make a small survey asking 6 randomly chosen people their age (older or younger than 60 years) and which of the breakfast cereals (Cornflakes, Frosties, Sugar Puffs, Branflakes) they like. Each respondent provides a vector with entries 1 or 0 corresponding to whether they like or dislike the cereal. Thus a respondent with (1101) would like Cornflakes, Frosties and Branflakes, but not Sugar Puffs. The older than 60 years respondents provide the following data (1000), (1001), (1111), (0001). The younger than 60 years old respondents responded (0110), (1110). A novel customer comes into the supermarket and says

she only likes Frosties and Sugar Puffs. Using naive Bayes trained with maximum likelihood, what is the probability that she is younger than 60? Work out the probability by hand, without using any code.

3. **Compact Naive Bayes Classification Using Matrices** (Don't worry, the linear algebra requirements here consist only of knowing how to multiply matrices!) Consider a naive Bayes classification problem with C classes, in which each data point has D categorical features, where feature d has K_d possible values. Let $\mathbf{X}$ be the $N \times (1 + \sum_d K_d)$ input matrix in binary "dummy variable form", where row n corresponds to data point n, the first column contains all 1s, the next K_1 columns contain exactly one 1 in each row, indicating which value feature d takes for each case; the next K_2 columns similarly indicate the value of feature 2; etc. Show that, given the class prior probabilities, $\pi_1, \ldots, \pi_C$, and the class conditional probabilities $\{\theta_{cdk}\}$, $c = 1, \ldots, C$, $d = 1, \ldots, D$, $k = 1, \ldots, K_d$, where

$$\theta_{cdk} := P(x_d = k \mid t = c)$$

there exists a matrix $\mathbf{A}$ such that the matrix product $\mathbf{XA}$ consists of unnormalized log posterior probabilities for the data. In the product $\mathbf{XA}$, the nth row corresponds to case n, and the cth column corresponds to the posterior probability of being in class c. Indicate what goes in each entry of $\mathbf{A}$. That is, the (n, c) entry of $\mathbf{XA}$ is

$$\log p(t_n = c \mid \mathbf{x}_n) + b$$

where b is some constant that does not depend on c (but may depend on $\mathbf{x}_n$). Hint: write both the desired unnormalized log posterior probability and the matrix product as sums; the latter consisting initially of variables, a_{ij}, and just match coefficients.

4. **Cancer Screening** The data for this problem consists of several diagnostic variables from tumors from each of 699 breast cancer patients (modified from a dataset in the University of California Irvine Machine Learning Repository[1]). The class variable is binary: Is the tumor malignant? The nine diagnostic variables are measurements of things like mean cell size, variability of cell sizes, various shape measures, etc. Each diagnostic variable has been coded on an integer scale ranging from 1 to 10. I have randomly divided the full dataset into training and test sets: `cancer_train.csv` and `cancer_test.csv`, containing 2/3 and 1/3 of the cases, respectively. The first column of the data consists of ID codes; the second is the

[1] `http://archive.ics.uci.edu/ml/`

binary `Malignant` label (0 or 1); and the remaining columns are the diagnostic features. Your mission (should you choose to accept it) is to design a Naive Bayes classifier that reports for a novel case a probability that it is malignant. Employ two variants of the classifier: one using maximum likelihood estimation to find the class conditional distributions and one using Bayesian estimation with a conjugate prior. Note that some of the cases have missing values for one of the features, `BareNuclei`. These missing values are denoted by -1 in the data. Note also that for several features, not all of the values 1-10 appear in both tumor types. You may choose to use the feature values as they are, or to bin them (since they consist of ordered values). If you choose to bin, you might select bins that have equal numbers of feature values, or bins that have approximately equal numbers of cases aggregated over classes. You may want to use cross-validation to select a binning scheme. Unlike in the digits problem on HW1, the classes are asymmetric, the misclassification risks are different, and the classifications are probabilistic, so you may want to use a more sophisticated metric than simple hard misclassification rate that takes these things into account. Do/report/answer each of the following.

(a) Explain the shortcomings of maximum likelihood estimation when it comes to the zero counts.

(b) Explain why naive Bayes classification provides a natural method for handling missing values, and why, on the other hand, missing data is a problem for k-nearest-neighbors classification.

(c) For the Bayesian method, use cross-validation to find the best parameters for the conjugate prior on the class conditional distributions. You can assume that the prior is symmetric over the feature values for each feature. It is possible that the optimal choices of concentration parameters are different for each feature, but you can fix them to be the same.

(d) Report the misclassification rate, precision, recall, F_1 score, and any custom error metrics you used, for your two final models (i.e., after all cross-validation is complete, and binning schemes and prior parameters are chosen): one based on MLE and one based on Bayesian estimation.

2. Naive Bayes With Continuous Features

5. For this problem we revisit the "5s and 9s" digits data that we classified on HW1 using K-nearest-neighbors. This time, to make things more manageable, as well as to make the Naive-Bayes assumption more reasonable, I have performed a dimensionality-reduction algorithm called Principal Components Analysis (PCA)

on the data, centering and rotating the data so that instead of representing individual pixel values, the features represent orthogonal linear combinations of pixel values; that is, I have multiplied the feature matrix of the original training set, $\mathbf{X}$, by another matrix $\mathbf{R}$ so that the columns of $\tilde{\mathbf{X}} := \mathbf{RX}$ are uncorrelated. Then, I sorted the columns in descending order according to their standard deviations, and retained only the first 11 columns (the idea being that most of the image is redundant and/or uninformative). I then applied the same rotation, column-permutation, and truncation to the test set (note that I did not use the test set in computing the rotation and permutation, so the columns of the test feature matrix will *not* be perfectly uncorrelated, and the standard deviations may not be in order). The resulting data files, `digits_reduced_train.csv` and `digits_reduced_test.csv`, consist of a column of either 5 or 9, indicating which digit it is, followed by the 11 input features.

(a) First create scatterplots of a few different pairs of features, color-coding by digit type. You should see two distinct but overlapping blobs.

(b) Let's first see how naive Bayes does using only the first two feature dimensions. Use diagonal bivariate Gaussians (i.e., two-dimensional Normal distributions in which the coordinates are uncorrelated) to model the class-conditional distributions (i.e., make the naive Bayes assumption that the class-conditional feature dimensions are independent), and find maximum likelihood estimates of the parameters (i.e., the means and variances of each class). Identify the parameters found.

(c) Use the resulting bivariate model to do naive Bayes classification of the test set. Report the misclassification rate.

(d) Use cross-validation to select the number of features you use (since the features are sorted in descending order of variability, you can just consider classifiers that use the first D features, rather than trying to explore all combinations). Report the misclassification rate for the best model on the test set.

(e) Repeat the above, this time using Gamma priors for the inverse variances (or, alternatively, inverse-Gamma priors for the variances — see the last problem for the PDF of an inverse-Gamma). To keep things simple(r)), use MLE for the mean and treat it as fixed. Do the classification using the (posterior) predictive distribution (i.e., integrate out the variances). Use the fact that if $(\sigma_d^2)^{-1}$ has a $\mathsf{Gamma}(a_{post}, b_{post})$ posterior and we use a Normal likelihood for $(x_d)_{new}$ with (known) mean μ and (unknown) variance σ_d^2, then the predictive

distribution for $(x_{new})_d$ is a scaled and shifted t-distribution, which has PDF:

$$p((x_{new})_d \mid a_{post}, b_{post}, \mu) = \frac{\Gamma(a_{post} + \frac{1}{2})}{\Gamma(a_{post})\sqrt{2\pi b_{post}}} \left(1 + \frac{(x_{new} - \mu)^2}{2b_{post}}\right)^{-(a_{post}+\frac{1}{2})}$$

Or if you prefer dealing with standardized distributions, you can define

$$(\tilde{x}_{new})_d := \frac{(x_{new})_d - \mu}{\sqrt{b_{post}/a_{post}}}$$

in which case

$$p((\tilde{x}_{new})_d \mid a_{new}, \mu) = \frac{\Gamma(a_{post} + \frac{1}{2})}{\Gamma(a_{post})\sqrt{2\pi a_{new}}} \left(1 + \frac{(\tilde{x}_{new})_d^2}{2a_{post}}\right)^{-(a_{post}+\frac{1}{2})}$$

(Note: You will need to find the appropriate posterior parameters, a_{post} and b_{post}, first before finding the predictive density! Use priors with high variance, e.g., $\mathsf{Gamma}(0.01, 0.01)$, for the $(\sigma_d^2)^{-1}$).

3. Bayesian Linear Regression

6. We are not quite ready to do a fully Bayesian treatment of the linear regression model – that is, to put priors on both $\mathbf{w}$ and σ^2. We will be able to apply this more realistic approach when we develop approximate inference methods. For now, assume that we are using a constant noise variance model with pre-specified σ^2, and that we put a $\mathcal{N}(0, \sigma_0^2\mathbf{I})$ prior on $\mathbf{w}$, with likelihood defined by

$$t_n \mid \mathbf{x}_n, \mathbf{w}, \sigma^2 \overset{ind.}{\sim} \mathcal{N}(\mathbf{x}_n\mathbf{w}, \sigma^2)$$

   (a) (Adapted from FCML 3.9) Show that the posterior mean for $\mathbf{w}$ is equivalent to the ridge regression solution:

$$\hat{\mathbf{w}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$$

   for some λ, and find that λ in terms of the parameters of the generative model and/or prior.

   (b) (Adapted from FCML 3.11) As the prior precision for $\mathbf{w}$ is increased (equivalently, the prior variance, σ_0^2 is decreased), we are saying that we are more and more certain going into data-collection that the regression coefficients should be small. Using the `womens100.csv` olympic dataset from HW1, calculate the log marginal likelihood (work with logs for numerical stability) for different polynomial orders and different choices of σ_0^2. Center and scale each column of the $\mathbf{X}$ polynomial matrix separately so that the equal variance prior makes

sense. How does the prior precision/variance impact the choice of polynomial order?

7. (Adapted from FCML 3.12) If we did put conjugate priors on both $\mathbf{w}$ and σ^2, then the prior on $\mathbf{w}$ would be Normal, and the prior on σ^2 would be inverse Gamma:

$$p(\sigma^2 \mid a, b) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-a-1} \exp\left\{-\frac{b}{\sigma^2}\right\}$$

and assuming prior independence, the joint prior on $(\mathbf{w}, \sigma^2)$ would be the product of Normal and inverse-gamma densities. Show that the joint posterior over $(\mathbf{w}, \sigma^2)$ is also the product of terms that look like $\mathcal{N}(\mu_{post}, \Sigma_{post})$ and inverse-Gamma(a_{post}, b_{post}) densities (albeit ones in which μ_{post} and Σ_{post} are interdependent; that is, the posterior density will not be separable into a term depending only on $\mathbf{w}$ and a term depending only on σ^2 — the weights and variance are no longer independent after conditioning on the data). Find the parameters (though, again, they will depend on each other).