

CS340 Winter 2010: HW3

Out Wed. 2nd February, due Friday 11th
February

1 PageRank

You are given in the file `adjency.mat` a matrix G of size $n \times n$ where $n = 1000$ such that

$$G_{i,j} = \begin{cases} 1 & \text{if outbound link from } i \text{ to } j, \\ 0 & \text{otherwise.} \end{cases}$$

You will have to construct the associated transition matrix

$$T_{i,j} = \begin{cases} p \times G_{i,j} / \left(\sum_j G_{i,j} \right) + (1-p) \times 1/n & \text{if } \sum_j G_{i,j} > 0 \\ 1/n & \text{otherwise} \end{cases}$$

where $p = 0.85$.

1. Implement the Power method in Matlab to compute the rank vector associated to T . Write also a Matlab script that plots

$$\log \|\mathbf{v}_k - \mathbf{v}_{k-1}\|$$

as a function of the iteration index $k = 1, 2, \dots, 100$. This allows us to see how the algorithm is converging and to determine how many iterations are necessary. Hand in the code and plot. Label the axes appropriately. Give the indexes of the three Webpages having the highest ranks.

2. Using

$$(T^T)^k \mathbf{v}_0 = a_1 \boldsymbol{\pi} + \sum_{i=2}^n a_i \lambda_i^k \mathbf{u}_i$$

where $\mathbf{u}_1 = \boldsymbol{\pi}, \mathbf{u}_2, \dots, \mathbf{u}_n$ are the orthonormal eigenvectors associated to the eigenvalues $\lambda_1 = 1 > |\lambda_2| \geq \dots \geq |\lambda_n|$ and the definition of \mathbf{v}_k , show that for k large enough

$$\|\mathbf{v}_k - \mathbf{v}_{k-1}\| \approx \left| \frac{a_2}{a_1} (\lambda_2 - 1) \right| |\lambda_2|^{k-1}.$$

Explain whether the plot obtained in part 1 is in agreement with this result. If it is not, provide an explanation for the possible discrepancy you have observed. Finally suggest a method to estimate λ_2 from the plot.

2 Analysis of A Naive Bayes Classifier

The first question asks you to analyse the following naive Bayes model that describes the weather in an imaginary country. We have

$$Y \in \{\text{night, day}\}, X_1 \in \{\text{cold, hot}\}, X_2 \in \{\text{rain, dry}\}$$

and

$$P(x_1, x_2, y) = P(y) P(x_1|y) P(x_2|y)$$

with $P(Y = \text{day}) = 0.5$, $P(X_1 = \text{hot} | Y = \text{day}) = 0.9$, $P(X_1 = \text{hot} | Y = \text{night}) = 0.2$, $P(X_2 = \text{dry} | Y = \text{day}) = 0.75$ and $P(X_2 = \text{dry} | Y = \text{night}) = 0.4$.

For each of the following formulae except the first, write an equation which defines it in terms of formulae that appear earlier in the list. (For example, you should give a formula for $P(x_1, x_2)$ in terms of $P(x_1, x_2, y)$). Then given the model above, calculate and write out the value of the formula for possible each combination of values of the variables that appear in it.

1. $P(x_1, x_2, y)$.
2. $P(x_1, x_2)$.
3. $P(y|x_1, x_2)$.
4. $P(x_1)$.
5. $P(x_2)$.
6. $P(x_1|x_2)$.

7. $P(x_2|x_1)$.
8. $P(x_1|x_2, y)$.
9. $P(x_2|x_1, y)$.

Are X_1 and X_2 conditionally independent given Y ? Are X_1 and X_2 marginally independent, integrating over Y ? Provide a short proof for both answers.

3 Bayes Classifier for Gaussian Data

Consider the following training set of heights x (in inches) and gender y (male/female) of some college students.

x	67	79	71	68	67	60
y	m	m	m	f	f	f

1. Fit a Bayes classifier to this data, i.e. estimate the parameters of the class conditional densities

$$p(x|y=c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{(x-\mu_c)^2}{2\sigma_c^2}\right)$$

and the class prior

$$p(y=c) = \pi_c$$

using Maximum Likelihood (ML) estimation for $c = m, f$. What are your ML estimate $\hat{\theta}$ of $\theta = (\mu_m, \sigma_m, \pi_m, \mu_f, \sigma_f, \pi_f)$?

2. Compute $p(y = m|x = 72, \hat{\theta})$.
3. What would be a simple way to extend this technique if you had multiple attributes per person, such as height and weight? Write down your proposed model as an equation.

4 Maximum Likelihood

The next question asks you to devise ML and Bayesian MAP estimators for a simple model of an uncalibrated sensor. Let the sensor output, X , be a random variable that ranges over the real numbers. We assume that, when tested over a range of environments, its outputs are uniformly distributed on some unknown interval $(0, \theta)$, so that

$$p(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{for } x \in (0, \theta) \\ 0 & \text{otherwise.} \end{cases}$$

We denote this distribution by $X \sim \text{Unif}(0, \theta)$. To characterize the sensor's sensitivity, we would like to infer θ .

1. Given N independent observations $x_{1:N} = (x_1, x_2, \dots, x_N)$ where $X_i \sim \text{Unif}(0, \theta)$, what is the likelihood function $L(\theta) = p(x_{1:N}|\theta)$? What is the maximum likelihood (ML) estimate of θ ? Give an informal proof that your estimator is in fact the ML estimator.
2. Suppose that we place the following prior distribution on θ :

$$p(\theta) = \alpha\beta^\alpha\theta^{-\alpha-1}\mathbb{I}_{(\beta, \infty)}(\theta)$$

This is known as a Pareto distribution. We denote it by $\theta \sim \text{Pareto}(\alpha, \beta)$. Plot the three prior probability densities corresponding to the following three hyperparameter choices: $(\alpha, \beta) = (0.1, 0.1)$, $(\alpha, \beta) = (2.0, 0.1)$ and $(\alpha, \beta) = (1.0, 1.0)$.

3. If $\theta \sim \text{Pareto}(\alpha, \beta)$ and we observe N independent observations $X_i \sim \text{Unif}(0, \theta)$, derive the posterior distribution $p(\theta|x_{1:N})$. Is this a member of any standard family?
4. For the posterior derived in part 3, what is the corresponding MAP estimate of θ ? How does this compare to the ML estimate?
5. Recall that the quadratic loss is defined as $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. For the posterior derived in part 3, what estimator $\hat{\theta}$ of θ minimizes the posterior expected quadratic loss? Simplify your answer as much as possible.

- Suppose that we observe three observations $x_{1:3} = (0.7, 1.3, 1.9)$. Determine the posterior distribution of θ for each of the priors (α, β) in part 2, and plot the corresponding posterior densities.

5 Naive Bayes Classifier Implementation

The Nursery database records a series of admission decisions to a nursery in Ljubljana, Slovenia. These data were downloaded from <http://archive.ics.uci.edu/ml/datasets/Nursery>. The database contains one tuple for each admission decision. The features or attributes include financial status of the parents, the number of other children in the house, etc. See <http://archive.ics.uci.edu/ml/machine-learning-databases/nursery/nursery.names> for more information about these features. The first three tuples in the dataset are as follows:

```
usual,proper,complete,1,convenient,convenient,nonprob,recommended,recommend
usual,proper,complete,1,convenient,convenient,nonprob,priority,priority
usual,proper,complete,1,convenient,convenient,nonprob,not_recom,not_recom
```

where the first 8 values are features or attributes and the 9th value is the class assigned (i.e., the admission decision recommendation).

Your job is to build a Naive Bayes classifier that will make admission recommendations. The data have been prepared for you by Marcos and can be downloaded from the course Webpage. All of the symbols have been replaced with identifying integers.

The first three rows of this matrix are:

```
>> load nursery.mat;
>> nursery(1:3,:)
ans =
1 1 1 1 1 1 1 1 2
1 1 1 1 1 1 1 1 2 4
1 1 1 1 1 1 1 1 3 1
```

You should divide this data into equal-sized training and testing data sets as follows (the `reset` ensures that we will all use the same training/test split).

```
load nursery.mat;
reset(RandStream.getDefaultStream)
nursery = nursery(randperm(size(nursery,1)),:);
train = nursery(1:size(nursery,1)/2,:);
```

```
test = nursery(size(nursery,1)/2+1:end,:);
```

1. Estimate a Naive Bayes model using Maximum Likelihood (ML) on **train** and use it to predict the y labels of the **test** data. To model the class conditional distribution of each feature x_i ($i = 1, \dots, 8$) and the class distribution, use multinomial/multinoulli distributions. Report the accuracy of your classifier, and submit your code.

2. Modify the Nursery data by duplicating the last attribute 20 more times. You can do this by executing

```
nursery = [nursery(:,1:end-1), repmat(nursery(:,end-1),1,20),nursery(:,end)];
```

before splitting into **train** and **test**. Then run your ML Naive Bayes estimator on the new training data, and evaluate it on the new testing data. Explain in words why you see the change in accuracy that you observe.

3. Using the original data and the model you estimated in part 1, calculate the (joint) log-likelihood $\sum_i \log P(\mathbf{x}^i, y^i)$ of the training data. Using this model, is it possible to calculate the (joint) log likelihood of the test data? Explain your answer.
4. Now use Bayesian MAP estimators using Dirichlet priors having all their hyperparameters set to $\alpha_k = 2$ to estimate all of the multinomial/multinoulli distributions parameters in your Naive Bayes model from the original training data. What accuracy do you obtain on the test data using this model? Calculate the (joint) likelihood of the training data under the MAP model. Is it higher or lower than the likelihood of the training data under the ML model? Explain your answer. Now calculate the likelihood of the test data under the MAP model. Explain why you don't run into the same problems.