CS440/ECE448 Spring 2018

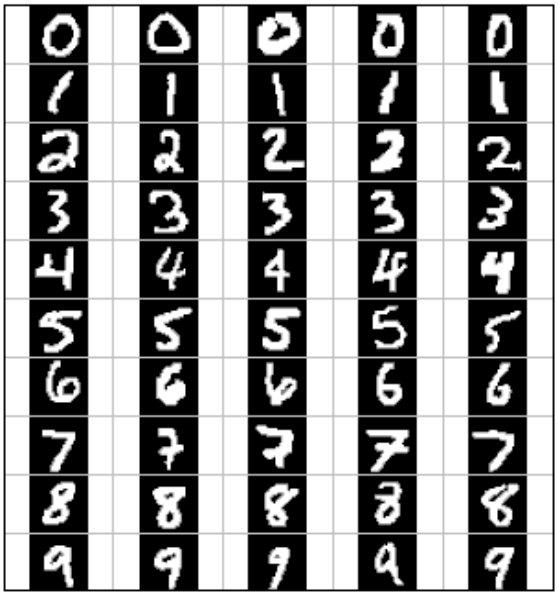
Assignment 3: Pattern Recognition

Due: Monday, April 9, 11:59:59PM

The goal of this assignment is to implement two different classifiers (Naive Bayes and Perceptrons), and apply both of the classifiers to the task of classifying visual patterns. As before, you can work in teams of up to three people (three-credit students with three-credit students, four-credit students with four-credit students).

Contents

Part 1: Naive Bayes Classifiers on Digit classification



(Adapted from Berkeley CS 188 project 5)

Data: This file is a zip archive containing training and test digits, together with their ground truth labels (see `readme.txt` in the zip archive for an explanation of the data format). There are roughly `2400` training exemplars (roughly `240` per class) and `400` test exemplars (roughly `40` per class).

Part 1.1 Single pixels as features (for everyone; 14 points)

- **Features:** The basic feature set consists of a single binary indicator feature for each pixel. Specifically, the feature F_{ij} indicates the status of the `(i,j)-th` pixel. Its value is `1` if the pixel is foreground, and `0` if it is background. The images are of size `32*32`, so there are `1024` features in total.

- **Training:** The goal of the training stage is to estimate the **likelihoods** $P(F_{ij} \mid class)$ for every pixel location `(i,j)` and for every digit class from `0` to `9`. The likelihood estimate is defined as

  $$P(F_{ij} = f \mid class) = (\#\text{ of times pixel (i,j) has value f in training tokens from this class}) / (\text{Total \# of training tokens from this class})$$

  In addition, as discussed in the lecture, you have to **smooth** the likelihoods to ensure that there are no zero counts. *Laplace smoothing* is a very simple method that increases the observation count of every value `f` by some constant `k`. This corresponds to adding `k` to the numerator above, and `k*V` to the denominator (where V is the number of possible values the feature can take on). The higher the value of k, the stronger the smoothing. Experiment with different values of k (say, from 0.1 to 10) and find the one that gives the highest classification accuracy.

  You should also estimate the **priors** $P(class)$ by the empirical frequencies of different classes in the training set.

- **Testing:** You will perform **maximum a posteriori (MAP)** classification of test digits according to the learned Naive Bayes model. Suppose a test image has feature values $f_{1,1}, f_{1,2}, \ldots, f_{32,32}$. According to this model, the posterior probability (up to scale) of each class given the digit is given by

  $$P(class) \cdot P(f_{1,1} \mid class) \cdot P(f_{1,2} \mid class) \cdot \ldots \cdot P(f_{32,32} \mid class)$$

  Note that in order to avoid underflow, it is standard to work with the log of the above quantity:

  $$\log P(class) + \log P(f_{1,1} \mid class) + \log P(f_{1,2} \mid class) + \ldots + \log P(f_{32,32} \mid class)$$

  After you compute the above decision function values for all ten classes for every test image, you will use them for MAP classification.

- **Evaluation:** Use the true class labels of the test images from the `testlabels` file to check the correctness of the estimated label for each test digit. Report your performance in terms of the **classification accuracy for each digit** (percentage of all test images of a given digit correctly classified). Also report your **confusion matrix**. This is a 10x10 matrix whose entry in row r and column c is the **percentage** of test images from class r that are classified as class c. In addition, for each digit class, show the test tokens from that class that have the highest and the lowest posterior probabilities according to your classifier. You can think of these as the most and least "prototypical" instances of each digit class (and the least "prototypical" one is probably misclassified).

  Important: The ground truth labels of test images should be used *only* to evaluate classification accuracy. They should not be used in any way during the decision process.

  Tip: You should be able to achieve at least 85% accuracy on the test set. One "warning sign" that you have a bug in your implementation is if some digit gets 100% or 0% classification accuracy (that is, your system either labels all the test images as the same class, or never wants to label any test images as some particular class).

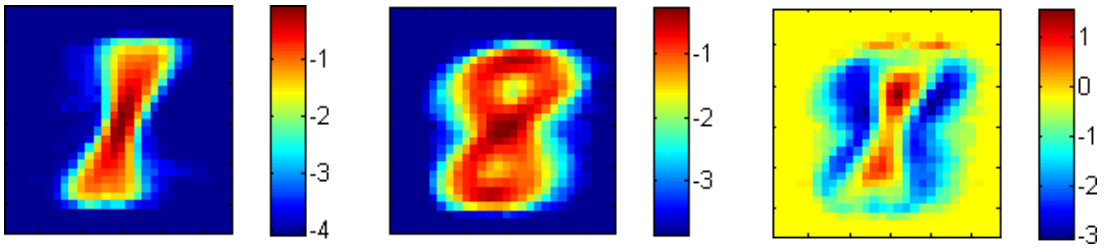
- **Odds ratios:** When using classifiers in real domains, it is important to be able to inspect what they have learned. One way to inspect a naive Bayes model is to look at the most likely features for a given label. Another tool for understanding the parameters is to look at odds ratios. For each pixel feature F_{ij} and pair of classes c_1, c_2, the odds ratio is defined as

  $$odds(F_{ij}=1, c_1, c_2) = P(F_{ij}=1 \mid c_1) / P(F_{ij}=1 \mid c_2).$$

  This ratio will be greater than one for features which cause belief in c_1 to increase over the belief in c_2. The features that have the greatest impact on classification are those with both a high probability (because they appear often in the data) and a high odds ratio (because they strongly bias one label versus another).

  Take four pairs of digit types that have the highest confusion rates according to your confusion matrix, and for each

pair, display the maps of feature likelihoods for both classes as well as the odds ratio for the two classes. For example, the figure below shows the log likelihood maps for 1 (left), 8 (center), and the log odds ratio for 1 over 8 (right):



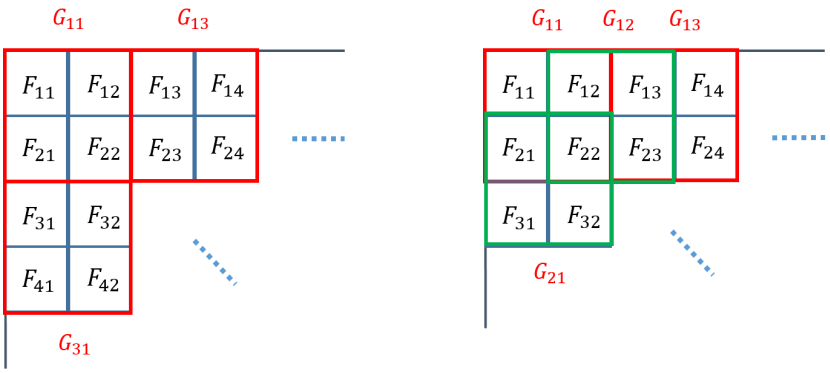
If you cannot do a graphical display like the one above, you can display the maps in ASCII format using some coding scheme of your choice. For example, for the odds ratio map, you can use '+' to denote features with positive log odds, ' ' for features with log odds close to 1, and '-' for features with negative log odds.

Part 1.2 Pixel groups as features (4 points required for 4-credit students; 2 extra credit points optional for 3-credit students)

Credit: Yanglei Song

Instead of each feature corresponding to a single pixel, we can form features from groups of adjacent pixels. We can view this as a relaxation of the Naive Bayes assumption that allows us to have a more accurate model of the dependencies between the individual random variables. Specifically, consider a 2*2 square of pixels with top left coordinate i,j and define a feature $G_{i,j}$ that corresponds to the ordered tuple of the four pixel values. For example, in the figure below, we have

$G_{1,1} = (F_{1,1}, F_{1,2}, F_{2,1}, F_{2,2})$.



(The exact ordering of the four pixel values is not important as long as it's consistent throughout your implementation.) Clearly, this feature can have 16 discrete values. The 2*2 squares can be disjoint (left side of figure) or overlapping (right side of figure). In the case of disjoint squares, there are 16 * 16 = 256 features; in the case of overlapping squares, there are 31 * 31 = 961 features.

We can generalize the above examples of 2*2 features to define features corresponding to n*m disjoint or overlapping pixel patches. An n*m feature will have 2^{n*m} distinct values, and as many entries in the conditional probability table for each class. Laplace smoothing applies to these features analogously as to the single pixel features.

In this part, you should build Naive Bayes classifiers for feature sets of n*m disjoint/overlapping pixel patches and report the following:

- Test set accuracies for disjoint patches of size 2*2, 2*4, 4*2, 4*4.
- Test set accuracies for overlapping patches of size 2*2, 2*4, 4*2, 4*4, 2*3, 3*2, 3*3.
- Discussion of the trends you have observed for the different feature sets (including single pixels), in particular, why certain features work better than others for this task.
- Brief discussion of running time for training and testing for the different feature sets (which ones are faster and why, and how does the running time scale with feature set size).

Tip: You should be able to achieve over 80% accuracy with your best feature set.

Part 1 Extra Credit (2 points each)

Extra credit tasks are worth up to two points each, up to a maximum of 25 percent (maximum extra credit = 6 points for 3-credit students, 8 points for 4-credit students).

- Apply your Naive Bayes classifier with various features to this face data. It is in a similar format to that of the digit data, and contains training and test images and binary labels, where 0 corresponds to 'non-face' and 1 corresponds to 'face'. The images themselves are higher-resolution than the digit images, and each pixel value is either '#', corresponding to an edge being found at that location, or ' ', corresponding to a non-edge pixel.

Part 2: Digit Classification using Discriminative Machine Learning Methods

Part 2.1 Digit Classification with Perceptrons (for everyone; 10 points)

Apply the multi-class (non-differentiable) perceptron learning rule from lecture to the digit classification problem from Part 1.1. As before, the basic feature set consists of a single binary indicator feature for each pixel. Specifically, the feature $F_{i,j}$ indicates the status of the (i,j)-th pixel. Its value is 1 if the pixel contains value 1, and 0 if it is 0. The images are of size 32*32, so there are 1024 features in total. For a multi-class perceptron, you need to learn a weight vector for each digit class. Each component of a weight vector corresponds to the weight of a pixel, which makes it of length either 1024 (without bias) or 1025 (with bias). You should report the following:

- Training curve: overall accuracy on the *training set* as a function of the epoch (i.e., complete pass through the training data). It's fine to show this in table form.
- Overall accuracy on the test set.
- Confusion matrix.

To get your results, you should tune the following parameters (it is not necessary to separately report results for multiple settings, only report which options you tried and which one worked the best):

- Learning rate decay function;
- Bias vs. no bias;
- Initialization of weights (zeros vs. random);
- Ordering of training examples (fixed vs. random);
- Number of epochs.

Part 2.2 Digit Classification with Nearest Neighbor (4 points required for 4-credit students; 2 extra credit points optional for 3-credit students)

Implement a k-nearest-neighbor classifier for the digit classification task in Part 2.1. You should play around with different choices of distance or similarity function to find what works the best. In the report, please discuss your choice of distance/similarity function, and give the overall accuracy on the test set *as a function of k* (for some reasonable range of k, from 1 to 25, you can describe this function with a table and discuss a general trend). For the best choice of k, give the confusion matrix. As a baseline, report the running time for a single query (classify a single instance in the test dataset) by using brute force. Discuss how you can optimize its performance. Finally, compare your nearest-neighbor accuracy to the accuracies you got with Naive Bayes and Perceptron.

Part 2 extra credit (2 points each)

Extra credit tasks are worth up to two points each, up to a maximum of 25 percent (maximum extra credit = 6 points for 3-credit students, 8 points for 4-credit students).

- For digits, it is possible to visualize the learned perceptron weights for each class as an image. Show some visualizations and discuss what they tell us about the model learned by the classifier -- in particular, which locations are the most important/discriminative for each digit. What do the signs of the weights tell us?
- Implement the differentiable perceptron learning rule and compare its behavior with the non-differentiable one.
- Apply any other classifier (support vector machine, decision tree, convolutional neural net, etc.) to digits, faces, or text. It is fine to use off-the-shelf code from an existing package (but be sure to cite your sources).

Report Checklist

Part 1:

1. For everybody:
    - Briefly discuss your implementation, especially the choice of the smoothing constant.
    - Report classification accuracy for each digit (note: this can be just the diagonal elements on the confusion matrix).
    - Show the confusion matrix.
    - For each digit, show the test tokens from that class that have the highest and lowest posterior probabilities according to your classifier.
    - Take four pairs of digit types that have the highest confusion rates, and for each pair, display feature likelihoods and odds ratio.
2. For four-credit students:
    - Report test set accuracies for disjoint patches of size 2*2, 2*4, 4*2, 4*4, and for overlapping patches of size 2*2, 2*4, 4*2, 4*4, 2*3, 3*2, 3*3.
    - Discuss trends for the different feature sets.
    - Discuss training and testing running time for different feature sets.

Part 2:

1. For everybody:
    - Discuss your implementation of perceptron classifier and parameter settings. Show your training curve, overall test-set accuracy, and confusion matrix.
2. For four-credit students:
    - Discuss your distance/similarity function, give the overall test set accuracy as a function of k, give confusion matrix for best k. Report running time for a single query by using brute force and discuss optimization measures. Compare your accuracy to Naive Bayes and perceptron.

Extra credit:

- We reserve the right to give **bonus points** for any advanced exploration or especially challenging or creative solutions that you implement. Three-credit students always get extra credit for submitting solutions to four-credit problems (at 50% discount). **If you submit any work for bonus points, be sure it is clearly indicated in your report.**

Statement of individual contribution:

- All group reports need to include a brief summary of which group member was responsible for which parts of the solution and submitted material. We reserve the right to contact group members individually to verify this

information.

WARNING: You will not get credit for any solutions that you have obtained, but not included in your report! For example, if your code prints out path cost and number of nodes expanded on each input, but you do not put down the actual numbers in your report, or if you include pictures/files of your output solutions in the zip file but not in your PDF, then **you will not get credit!**. The only exception is animated paths (videos or animated gifs).

Submission Instructions

As before, **one designated person from the group** will need to submit on **Compass 2g** by the deadline. Three-credit students must upload under **Assignment 3 (three credits)** and four-credit students must upload under **Assignment 3 (four credits)**. Each submission must consist of the following two attachments:

1. A **report** in **PDF format**. As before, the report should briefly describe your implemented solution and fully answer all the questions posed above. **Remember: you will not get credit for any solutions you have obtained, but not included in the report.**

   All group reports need to include a brief **statement of individual contribution**, i.e., which group member was responsible for which parts of the solution and submitted material.

   The name of the report file should be **lastname_firstname_MP3.pdf**. Don't forget to include the names of all group members and the number of credit credits at the top of the report.

2. Your **source code** compressed to a **single ZIP file**. The code should be well commented, and it should be easy to see the correspondence between what's in the code and what's in the report. You don't need to include executables or various supporting files (e.g., utility libraries) whose content is irrelevant to the assignment. If we find it necessary to run your code in order to evaluate your solution, we will get in touch with you.

   The name of the code archive should be **lastname_firstname_MP3.zip**.

Multiple attempts will be allowed but in most circumstances, only the last submission will be graded. **We reserve the right to take off points for not following directions.**

Late policy: For every day that your assignment is late, your score gets multiplied by 0.75. The penalty gets saturated after four days, that is, you can still get up to about 32% of the original points by turning in the assignment at all.

Be sure to also refer to **course policies** on academic integrity, etc.