10-701 Introduction to Machine Learning

Homework 1

Due Oct 1, submission instructions would be announced next week

---

Rules:

1. Homework is due on the due date in the class on October 1. Please see course website for policy on late submission.
2. We recommend that you typeset your homework using appropriate software such as LaTeX. If you are writing please make sure your homework is cleanly written and legible. The TAs will not invest undue effort to decrypt bad handwriting.
3. You are allowed to collaborate on the homework, but you should write up your own solution and code. Please indicate your collaborators in your submission.
4. The submission procedure will be updated and announced soon.

---

1 Probability and Statistics Review (25 Points) (Mrinmaya)

1.1 Exponential Families

Many commonly used distributions in Statistics and Machine Learning fall under the category of Exponential Family of distributions. Exponential family is a set of probability distributions whose probability density function (or probability mass function, if discrete) can be expressed in the form: $f_X(x) = h(x)\exp\left(\eta(\theta)\cdot T(x) - A(\theta)\right)$ where $T(x)$, $h(x)$, $\eta(\theta)$, and $A(\theta)$ are known.

(a) Show that Multinomial distribution, Multi-variate Gaussian distribution and Dirichlet distribution are members of the exponential family.

(b) Consider dataset $D = \{x_i\}_{i=1}^N$ which is independently and identically distributed (i.i.d.) according to some known exponential family distribution $f(x|\boldsymbol{\theta}) = h(x)\exp\left(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{T}(x) - A(\boldsymbol{\theta})\right)$. Let the prior for the parameter $\boldsymbol{\theta}$ be given by $p_\pi(\boldsymbol{\theta}|\boldsymbol{\chi},\nu) = f(\boldsymbol{\chi},\nu)\exp\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\chi} - \nu A(\boldsymbol{\theta})\right)$. Compute the posterior and show that it takes the same form as the prior.

Note: This notion is called "Conjugacy" and this will be very useful in Bayesian modelling.

1.2 Maximum Likelihood Estimation

We learnt about Maximum Likelihood estimation in class. For a fixed set of data and underlying statistical model, the method of maximum likelihood selects the set of values of the model parameters that maximises the likelihood function.

(a) Consider the model where the data $D = \{x_i\}_{i=1}^N$ is i.i.d. according to some known exponential family distribution. Further, let $\eta(\theta) = \theta$. Show that the maximum likelihood solution is given by: $A'(\hat{\theta}_{ML}) = \frac{1}{N}\sum_i T(x_i)$

(b) Now, consider the special case where the data is i.i.d according to the uniform distribution: $x_i \sim \mathit{uniform}(0,\theta)$ i.e. $p(x_i|\theta) = \begin{cases} \frac{1}{\theta} & x_i \in [0,\theta] \\ 0 & \mathit{otherwise} \end{cases}$

Now, compute the maximum likelihood estimator $\hat{\theta}_{ML}$.

2 Decision Boundary of Naive Bayes (10 Points) (Yuntian)

Consider the problem of predicting a label $Y \in \mathcal{Y}$ given an input feature vector $X \in \mathcal{X}$. Suppose $\mathcal{X} = \mathbb{R}^d$, $Y \in \{0,1\}$. We make the Naive Bayes assumption that features are conditionally independent given label, i.e. $P(X|Y) = \prod_{i=1}^d P(X_i|Y)$, where X_i denotes the i-th component of X. In addition, we impose a Bernoulli prior on Y: $P(Y=1) = \pi$.

(a) Assume that $P(X_i|Y)$ is in the exponential family: $P(X_i = x_i|Y = y) = h_i(x_i)\exp(\theta_{iy}\cdot T_i(x_i) - A_i(\theta_{iy}))$. Compute the posterior distribution $P(Y|X)$. Compute the decision boundary $\{x \in \mathcal{X} : P(Y = 1|X = x) = P(Y = 0|X = x)\}$. (You can use sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$ to simplify the expression).

(b) Assume that $P(X_i|Y = y)$ is a Gaussian distribution with mean μ_{iy} and variance σ_i^2 (Note that the variance here does not depend on label Y), show that the decision boundary is linear in terms of X. (Hint: Recall from Problem 1.1 that Gaussian distribution is in the exponential family).

3 KNN Classification (15 Points) (Yuntian)

Consider a classification problem using kNN. We have N training points $x_1, x_2, \cdots x_N$ and corresponding labels $y_1, y_2, \cdots y_N$. If we wish to classify a new data point, the classification rule is simply a majority vote among its k nearest neighbours in the training set.

(a) Consider $k = 1$ case, is it possible to build a decision tree (the decision at each node can only take the form of "$x^i \leq t$ or $x^i > t$" where x^i denotes the i-th feature of x, t is a real number and can be different in different nodes) which behaves exactly the same as the 1-NN classifier?

(b) Recall from class that the decision rule of kNN can be viewed using a probabilistic interpretation: Suppose we have a data set comprising of N_k points in class C_k. For a new point x, we draw a ball around x containing exactly K nearest neighbours. Suppose this ball has volume V and contains K_k points from class C_k. Then we can estimate the density conditioned on class C_k as:

$$P(x|C_k) = \frac{K_k}{N_k V}$$

Similarly, we estimate the unconditioned density as:

$$P(x) = \frac{K}{NV}$$

We also estimate the prior as:

$$P(C_k) = \frac{N_k}{N}$$

In class we have shown that under this model, Bayes decision rule will yield kNN classification. Show that the density $P(x)$ given by this model is an improper distribution whose integral over all space is not guaranteed to sum up to 1. (You only need to construct a special case to show that).

(c) In kNN classification, the nearest neighbours of a given test point depend on the distance metric. We often implicitly use Euclidean distance as the distance metric, but we can use other distance metrics as well. Consider a binary classification problem using 1-NN. Assume the labels can only be C_1 and C_2. Denote the number of training points as N. For a test point x, denote its nearest neighbour as x' (note that the nearest neighbour depends on the distance metric), then the probability that x is misclassified is:

$$P_N(e|x) = P(C_1|x)P(C_2|x') + P(C_2|x)P(C_1|x')$$

The asymptotic error rate (i.e. the error rate when the size of training set is infinite) is:

$$P(e|x) = \lim_{N\to\infty} P_N(e|x)$$

It can be shown that when $N \to \infty$, the asymptotic error rate $P(e|x)$ of 1-NN is upper bounded by twice the minimum achievable error rate. However, what we care about is $P_N(e|x)$. Therefore, if we can bound the difference between $P_N(e|x)$ and $P(e|x)$, then $P_N(e|x)$ is also upper bounded (As $P_N(e|x)$ depends on the training set, we can only bound it in a probabilistic sense, but that's not the focus of this problem). Our objective here is to use a proper distance metric to minimize the expected squared difference between $P(e|x)$ and $P_N(e|x)$: $\min \mathbb{E}[(P_N(e|x) - P(e|x))^2|x]$. Note that the expectation here is taken with respect to the training set, conditioned on the test point x.

Denote $\nabla P(C_1|x) \equiv \frac{\partial P(C_1|x)}{\partial x}$. Approximate $P(C_1|x')$ by the first order Taylor expansion at x: $P(C_1|x') \simeq P(C_1|x) + \nabla P(C_1|x)^T(x' - x)$, show that $\mathbb{E}[(P_N(e|x) - P(e|x))^2|x]$ is minimized by using the distance metric $d(x, x') \equiv |\nabla P(C_1|x)^T(x - x')|$.

Note: In practice, we can estimate the direction of $\nabla P(C_1|x)$ from training data. Denote it as $\hat{\nabla}$, then we can use $d(x, x') \equiv |\hat{\nabla}^T(x - x')|$ as the distance metric, because the scaling of $\nabla P(C_1|x)$ does not affect the nearest neighbour.

4 Decision Trees (25 Points) (Yuntian)

4.1 Decision tree on two features

Consider training a decision tree on n two dimensional vectors $x = (x_1, x_2)$.

(a) Assume we have two equal vectors x and x' in our training set (that is, all attributes of x and x' including the labels are exactly the same). Can removing x' from our training data change the decision tree we learn for this dataset? Explain briefly.

(b) Assume that the training instances are linearly separable. That is, there exists a $\{w, b\}$ such that

$$y = \begin{cases} +1 & w^T x + b > 0 \\ -1 & w^T x + b \leq 0 \end{cases}$$

Can a decision tree correctly classify these vectors? (the decision at each node can only take the form of "$x_i \leq t$ or $x_i > t$" where x_i is a feature, t is a real number and can be different in different nodes). If so, what is an upper bound on the depth of the corresponding decision tree (as tight as possible)? If not, why not?

(c) Now assume that these n inputs are not linearly separable (that is, no $\{w, b\}$ exists for correctly classifying all inputs using the above rule). Can a decision tree correctly classify these vectors? (the decision at each node can only take the form of "$x_i \leq t$ or $x_i > t$" where x_i is a feature, t is a real number and can be different in different nodes) If so, what is an upper bound on the depth of the corresponding decision tree (as tight as possible)? If not, why not?

4.2 C4.5 Algorithm

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sun	26	High	Low	No
D2	Sun	25	High	High	No
D3	Overcast	25	High	Low	Yes
D4	Rain	24	High	Low	Yes
D5	Rain	19	Normal	Low	Yes
D6	Rain	20	Normal	High	No
D7	Overcast	20	Normal	High	Yes
D8	Sun	23	High	Low	No
D9	Sun	20	Normal	Low	Yes
D10	Rain	25	Normal	Low	Yes
D11	Sun	24	Normal	High	Yes
D12	Overcast	22	High	High	Yes
D13	Overcast	23	Normal	Low	Yes
D14	Rain	23	High	High	No

Table 1: Dataset

We have learnt ID3 algorithm in class. One limitation of ID3 is that it is overly sensitive to features with large numbers of values. For example, if each training instance has a unique id, then the information gain will be maximized by using this id as feature, while this is not useful. C4.5 algorithm improves this by using information gain ratio instead of information gain to evaluate features. Denote the feature for an input by X, and its label by Y. Recall that in ID3, we choose feature X that maximizes information gain $IG(X)$:

$$IG(X) \equiv H(Y) - H(Y|X)$$

Now we define split information as follows: Suppose there are$|D|$ samples at the current node, and after splitting by feature X, they fall into V child nodes. Assume the number of samples that pass through each child node is $|D_1|, |D_2|, \cdots |D_V|$, then the split information of feature X is

$$SplitInfo(X) \equiv -\sum_{j=1}^{V} \frac{|D_j|}{|D|} \log \frac{|D_j|}{|D|}$$

The information gain ratio is defined as

$$GainRatio(X) \equiv \frac{IG(X)}{SplitInfo(X)}$$

C4.5 algorithm uses information gain ratio to determine the best splitting attribute. The intuition is that $SplitInfo(X)$ acts as a normalizer to penalize features with a large number of values. For example, if $\forall i, j$ we have $|D_i| = |D_j|$, then $SplitInfo(X) = \log V$, so features with smaller V is preferred.

Draw the decision tree of the dataset in table 1[1] using both the ID3 and C4.5 algorithms. The features here are Outlook, Temperature, Humidity and Wind, and the label to be predicted is Play. Treat Temperature as a discrete feature.

5 Naive Bayes for sentiment classification (25 Points) (Mrinmaya)

The goal of this assignment is to implement a Naive Bayes classifier as described in the lecture and to apply it to the task of sentiment classification. You are free to design the code as you wish and to implement the code in any language that you wish.

We will work with the following movie review data set: http://www.cs.cornell.edu/people/pabo/movie-review-data/review_polarity.tar.gz

The data set contains movie reviews classified as positive or negative reviews. We will provide standard train and test splits in the data - use reviews 000 to 799 for training and reviews 800-999 (in both pos and neg classes) for testing. Ignore the cross-validation tags. Please use this exact split only. There are two steps to this assignment: the pre-processing step and the classification step.

5.1 Pre-processing step

This first step converts the movie reviews into features to be used by the naive Bayes classifier. You will be using the bag of words approach described in class. For completeness, the following steps outline the process involved:

1. Form the vocabulary: Use all the positive and negative sentiment carrying words provided by Hu and Liu: http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar[2] as the vocabulary. The words in the vocabulary will form the feature set. You may want to keep the vocabulary in alphabetical order to help you with debugging your assignment, but this is not necessary.

2. Now, convert the training data into a set of features. Let M be the size of your vocabulary. For each review, you will convert it into a feature vector of size M. Each slot in that feature vector takes the value of 0 or 1. If the i^{th} slot is 1, it means that the i^{th} word in the vocabulary is present in the review; otherwise, if it is 0, then the i^{th} word is not present in the review. Most of the feature vector slots will be 0 so it will be efficient if you only store the indices that have 1's.

5.2 Classification step

Build a naive Bayes classifier as described in class.

1. In the first phase, which is the training phase, the naive Bayes classifier reads in the training data along with the training labels and learns the parameters used by the classifier.

[1] http://saiconference.com/Downloads/SpecialIssueNo10/Paper_3-A_comparative_study_of_decision_tree_ID3_and_C4.5.pdf

[2] You will need to remove the header information lines from the files positive-words.txt and negative-words.txt

2. In the testing phase, the trained naive Bayes classifier classifies the data in the testing data file. You will need to convert the reviews in the testing data into a feature vector, just like in the training data where a 1 in the i^{th} slot indicates the presence of the i^{th} word in the vocabulary while a 0 indicates the absence. If you encounter a word in the testing data that is not present in your vocabulary, ignore that word.

3. Output the accuracy of the naive Bayes classifier by comparing the predicted class label of each review in the testing data to the actual class label. The accuracy is the number of correct predictions divided by the total number of predictions.

Note: Make sure that you follow the implementation hints given in the lecture. Specifically, you may want to do the probability calculations in log space to prevent numerical instability. Also, use laplace prior of 0.1 (also called "add 0.1 smoothing") to avoid zero probabilities.

5.3 Results

Your results must be reported in a write-up (to be attached with the written component of this assignment).

(a) Train Accuracy: Run your classifier by training on the training split and then testing on training split itself. Report the accuracy. In this situation, you are training and testing on the same data. This is a sanity check: your accuracy should be fairly high.

(b) Test Accuracy: Run your classifier by training on the training split and then testing on testing split. Report the accuracy. Are you over-fitting?

(c) Most sentiment carrying words: For both positive and negative classes, print top 10 words i.e. words that have highest $p(word|sentiment = positive)$ and $p(word|sentiment = negative)$.

5.4 Experiment with your code

Now, we will play around with the code a little bit and try to boost up the classifier accuracy if we can.

Negation Handling: A major problem faced during the task of sentiment classification is that of handling negations. Since we are using each word as feature, the word "good" in the phrase "not good" will be contributing to positive sentiment rather that negative sentiment as the presence of "not" before it is not taken into account. To mitigate this, we will introduce a simple fix - for all words preceded by a not or n't in the training set, introduce a new feature called "not_" + word. Add these features to the ones included in the vocabulary you used above.

Perform negation handling and report Train and Test Accuracy on the data set as before. Are you over-fitting? Does your accuracy increase? Why or Why not? Explain in your report.

Including Bi-grams: Often, information about sentiment is conveyed by adjectives or more specifically by certain combinations of adjectives with other parts of speech. This information can be captured by adding features like consecutive pairs of words (also called bi-grams) where the first word is an adverb of degree and the second word is a sentiment carrying word. Words like "very" or "definitely" don't provide much sentiment information on their own, but phrases like "very bad" or "definitely recommended" increase the probability of a document being negatively or positively biased. To utilize this we will add some more bi-grams to our vocabulary. Specifically, we will add all pairs where the first word is one of the following seven adverbs: "extremely", "quite", "just", "almost", "very", "too", "enough", and the second word is any word in the original vocabulary you used above.

Perform bi-gram inclusion and report Train and Test Accuracy on the data set as before. Are you over-fitting? Does your accuracy increase? Why or Why not? Explain in your report.

5.5 Extra Credit (5 Points)

Now that you have your cool Sentiment classifier ready, try to improve it a little more and earn a bonus of 5 more points.

One idea to do this is to allow a vocabulary of all words in your training set (do not use the Hu and Liu vocabulary), perform negation handling and include all bi-grams (all consecutive words in the training set) to your vocabulary. The use of higher dimensional features like all possible bi-grams will result in a blow up in the number of features. So it might be useful to down select our features. While we have not covered feature selection in class, you are encouraged to read up on these and implement one. Our suggestion is to do "Forward Greedy Selection" but you are free to implement one of your choice.

Describe your feature selector. How many features did you finally select? Why? Report your Train and Test Accuracy on the data set as before. Are you over-fitting now? Does your accuracy increase? Why or Why not? Explain in your report.

5.6 Submission Requirements

You are required to submit your code as well as your report. Attach your report with the written component of the assignment. Submission instructions will be announced soon.