CMPS 242: Machine Learning: Fall 2016

Mid-Term Practice Qns

Question 1 (True-False Questions):

The following questions should be answered as *true* or *false*. If you answer true, provide a short justification, if false explain why or provide a small counter-example.

(1) A classifier trained on less training data is less likely to overfit.

(2) The decision boundary obtained by a Logistic Regression classifier is always linear.

(3) We can get multiple local optimum solutions if we solve a linear regression problem by minimizing the sum of squared errors using gradient descent.

(4) Discriminative algorithms allow you to classify points, without providing a model of how the points are actually generated.

(5) Logistic Regression and Linear Regression can both be used for curve fitting.

(6) Decision Rules for classifiers based on Bayes Rule such as Naive Bayes have the ability to automatically handle imbalanced classes.

Question 2 (Short Questions):

Answer the below questions as succinctly as possible.

(1) We estimate the probability of landing heads θ of a coin from the results of N flips. We use psuedo-counts to influence the "fairness" of the coin. This is equivalent to using which distribution as a prior for θ.

(2) Logistic Regression is named after the log-odds of success defined as:

$$\log\left(\frac{p(Y=1|X=x)}{p(Y=0|X=x)}\right)$$

Show that log-odds of success is a linear function of x.

(3) $X = (X_1, X_2)$ is drawn from a two dimensional Gaussian distribution with a diagonal covariance matrix.

$$X = (X_1, X_2) \sim \mathcal{N}(\mu, \Sigma) \tag{1}$$

$$\Sigma = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \tag{2}$$

where a and b are some real numbers. Are X_1 and X_2 independent? Explain as concisely as possible.

(4) Assume a dataset consisting of N data points $x = \{x_1, x_2, \ldots, x_N\}$, where each data point x_i has D features $\{x_{ij}\}, j = 1, \ldots, D$. A Naive Bayes classifier assumes that: (pick from one of the options listed below)

   (a) data points x_i are independent of each other
   (b) features x_{ij} are uncorrelated (independent) of each other
   (c) both (a) and (b)
   (d) neither (a) nor (b)

(5) Suppose we wish to calculate $p(H|E_1, E_2)$ and we have no conditional independence information. Which of the following sets of numbers are sufficient for the calculation?

   (a) $p(E_1, E_2),\ p(H),\ p(E_1|H),\ p(E_2|H)$
   (b) $p(E_1, E_2),\ p(H),\ p(E_1, E_2|H)$
   (c) $p(H),\ p(E_P1|H),\ p(E_2|H)$

   Suppose we know that $p(E_1|H, E_2) = p(E_1|H)$ for all values of H, E_1, E_2. Now which of the above three sets are sufficient?

(6) You have received a shiny new coin and want to estimate the probability θ that it will come up heads if you flip it. A priori you assume that the most probable value of θ is 0.5. You then flip the coin 3 times, and it comes up heads twice. Which will be higher, your maximum likelihood estimate (MLE) of θ, or your maximum a posteriori probability (MAP) estimate of θ? Explain intuitively why, *using only one sentence*.

(7) Assume you attempt to apply the techniques learnt in the class to do stock-market prediction. What is one assumption that you will find violated by the data you see?

(8) Provide one advantage and one disadvantage of using Maximum Likelihood method (MLE) for estimation.

(9) Suppose we have a dataset consisting of n data points, $x = \{x_1, \ldots, x_n\}$ assumed to originate from a gaussian distribution with parameter θ. Also assume each data point belongs to class t ($t = 1, \ldots, k$). Which of the following statements is False?

(a) A Naive Bayes classifier computes the probability $p(x|t, \theta)$

(b) A Naive Bayes classifier computes the probability $p(t|x, \theta)$

(c) A Logistic Regression classifier computes the probability $p(x|t, \theta)$

(d) A Logistic Regression classifier computes the probability $p(t|x, \theta)$

Question 3 (Long Questions):

(1) Suppose that a study shows that 80% of people who have contracted Zika disease (a type of mosquito borne disease) got bit by mosquitos prior to contracting the disease. Zika disease is incredibly rare; suppose only one in five million people have the disease.

- If you got bit by a mosquito, should you be worried?
- Does this depend on how many other people got bit my mosquitoes?

(Assume the probability that a random person gets bit by a mosquito to be $\frac{1}{10}$).

(2) Derive analytic expression for the posterior distribution when the likelihood is considered gaussian and priors are placed on both the mean and variance (for simplicity, consider only the univariate case, i.e: number of dimensions = 1).

(3) Suppose we have two dimensional data with features $x = (x_1, x_2)$. The two class-conditional densities are $p(x|\mathcal{C} = 1)$ and $p(x|\mathcal{C} = 2)$, are 2D Gaussian distributions centered at points $\boldsymbol{\mu}_1 = (4, 11)$ and $\boldsymbol{\mu}_1 = (10, 3)$ respectively with the same covariance matrix $\boldsymbol{\Sigma} = 2\mathbf{I}$. Suppose the priors are $P(\mathcal{C} = 1) = 0.6$ and $P(\mathcal{C} = 2) = 0.4$.

- Suppose we use a Bayes decision rule, derive the two posteriors.
- Derive the MAP decision rule for this problem in terms of the two features x_1 and x_2.
- Is the decision boundary for naive bayes classifier for this problem linear or quadratic? Provide a one sentence justification for either answer.

(4) Suppose we have a dataset of n emails, with each data point representing binary features of three words $w_1 =$ `birthday`, $w_2 =$ `cake`,

w_3 = `party` respectively depending upon if the corresponding word occurred in the email or not. Therefore, each data point is represented as a bit-string of length 3.

- Assume w_1, w_2, w_3 are correlated somehow. How many parameters do you need to estimate in order to compute the probability of a particular data point in this dataset, say $p(w_1 = 0, w_2 = 1, w_3 = 1)$? Do you see any challenges?
- What is a simplifying assumption you can make to compute the above probability much more easily? How many parameters do you now need to estimate after making this assumption?
- Write down an expression for the likelihood of this dataset.

(5) These two homework sets have good problems.

(a) `http://www.ee.columbia.edu/~sfchang/course/spr/handout/hw1.pdf`

(b) `http://www.ee.columbia.edu/~sfchang/course/spr/handout/hw2.pdf`