COMP 652: Machine Learning - Midterm exam

Sample Questions with Solutions
Posted March 5, 2015

1. **Maximum likelihood**

   Consider the following probability distribution:

   $$P_\theta(x) = 2\theta x e^{-\theta x^2}$$

   where θ is a parameter and x is a positive real number. Suppose you get m i.i.d. samples x_i drawn from this distribution. Show how one can compute the maximum likelihood estimator for θ based on these samples.

   Solution: We with down the likelihood under the iid assumption:

   $$L(D, \theta) = \prod_{i=1}^{m} P_\theta(x_i)$$

   Taking the log, we geT:

   $$\log L(D, \theta) = \sum_{i=1}^{m} \log P_\theta(x_i) = \sum_{i=1}^{m} (\log 2 + \log \theta + \log x_i - \theta x_i^2)$$

   Taking the derivative wrt θ, we get:

   $$\frac{\partial \log L(D, \theta)}{\partial \theta} = \sum_{i=1}^{m} \left(\frac{1}{\theta} - x_i^2 \right) = \frac{m}{\theta} - \sum_{i=1}^{m} x_i^2$$

   Setting this to 0 and solving for θ we get:

   $$\theta = \frac{m}{\sum_{i=1}^{m} x_i^2}$$

2. **One-sided error function**

   Suppose that you have a regression problem for which you have an error function of the following type:

   $$J(\mathbf{w}) = \begin{cases} h_{\mathbf{w}}(\mathbf{x}) - y & \text{if } h_{\mathbf{w}}(\mathbf{x}) - y > 0 \\ 0 & \text{otherwise} \end{cases}$$

   (a) Is this error function differentiable?

   Solution: It is differentiable everywhere except at 0 (there it is only continuous).

(b) Describe a learning algorithm (similar to an algorithm we discussed in class) that you can use to optimize this error function, assuming that $h_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$.

Solution: When $h_{\mathbf{w}}(\mathbf{x}) - y > 0$ we do gradient descent, so:

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha(h_{\mathbf{w}}(\mathbf{x}) - y)\mathbf{x}$$

If $h_{\mathbf{w}}(\mathbf{x}) - y \leq 0$, there is no update for $\mathbf{w}$.

3. **Kernels**

Suppose you are given a set of instances which are *directed graphs* with a known, fixed set of nodes V. The graphs differ in their set of edges. Consider a kernel function over two graphs G and G', computed as follows:

$$K(G, G') = \sum_{(a,b)\in V\times V} k((a, b))$$

where

$$k((a, b)) = \begin{cases} 1 & \text{if } (a, b) \in E(G) \text{ and } (a, b) \in E(G') \\ -1 & \text{if } (a, b) \in E(G) \text{ and } (b, a) \in E(G') \text{ or vice versa} \\ 0 & \text{otherwise} \end{cases}$$

where $E(G)$ and $E(G')$ are the sets of edges of the graphs. Assume that for any pair of nodes $(a, b) \in V \times V$, at most one of the edges (a, b) and (b, a) can be present in a graph (but not both).

(a) Is K a kernel? Justify your answer.

Solution: Without loss of generality, pick an ordering among the pairs of vertices from V. Consider a feature vector ϕ which has an entry for any pair of vertices $(i, j), i \leq j$, such that:

$$\phi_{(i,j)}(G) = \begin{cases} 1 & \text{if } (i, j) \in E(G) \\ -1 & \text{if } (j, i) \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

It is easy to see that $K(G, G') = \phi(G) \cdot \phi(G')$ so using the definition, K is a kernel.

(b) Suppose you replace the -1 value above with a real number q. For what range of q is K a kernel?

Solution: It works for -1 and 0 (the latter with a different feature vector).

4. **Maximum likelihood**

Consider the following probability density function:

$$P_\theta(x) = \frac{\theta e^{-x}}{(1 + e^{-x})^{(\theta+1)}}$$

where θ is a parameter in $(0, \infty)$ and x is a real number. Suppose you get m i.i.d. samples x_i drawn from this distribution. Show how one can compute the maximum likelihood estimator for θ based on these samples.

Solution: As before, the log likelihood under the iid assumption is:

$$\log L(D,\theta) = \sum_{i=1}^{m} \left(\log\theta - x_i - (\theta+1)\log(1+e^{-x_i})\right)$$

Taking the derivative wrt theta:

$$\frac{\partial \log L(D,\theta)}{\partial\theta} = \sum_{i=1}^{m}\left(\frac{1}{\theta} - \log(1+e^{-x_i})\right) = \frac{m}{\theta} - \sum_{i=1}^{m}\log(1+e^{-x_i})$$

Setting to 0 we get:

$$\theta = \frac{m}{\sum_{i=1}^{m}\log(1+e^{-x_i})}$$

5. **Bayes nets**

Suppose you have to analyze a medical data set, with 1 binary attribute (gender) and one numerical attribute (the result of a test). The outcome to predict is healthy or sick (binary classification). You use the following graph structure: O is the root, and G and T are conditionally independent given O, and both depend on O (this is also called a Naive Bayes structure).

(a) Suppose you will model the binary attribute using a binomial and the continuous one using a Gaussian distribution. How many parameters will the classifier have, and what do they represent?

Solution: We will have one parameter for the class node (the probability of the value being 1) and two parameters, μ and σ, for the continuous attribute, for each of the values of the class. For the binary attribute, you will have one parameter, describing the binomial, for each value of the class. So, in all 6 parameters.

(b) Draw a picture, with the two variables on the two axes, of a data set that would be well modelled by such a classifier. Points of the two classes would be $+$ and $-$ in this 2D space.

Solution: You will have lines on this graph, at the two values of the binary attribute, with each line representing the Gaussian. They should be such that the lines for the two classes are away from each other.

(c) Alternatively, you consider fitting two separate classifiers, one for male, one for female patients. Each one would be a Gaussian Naive Bayes classifier based on just the test result. How many parameters will need to be learned in this case, and what are they? How does the second solution compare to the first one in terms of bias and variance? Justify your answer.

Solution: Each classifier will have 5 parameters (one for the class, and two means and variances). At the same time, there is less data to train each classifier (since only patients of one gender are used). So, this alternative is lower-bias higher-variance. One can see the low bias by noting that in this case, the gender is not conditionally independent of the test result give the class, and different incidence of the disease is permitted based on gender (whereas this was not accommodated in the previous setting).

(d) Suppose that instead of this structure, you choose a v-structure, in which G and O are independent and T is dependent on both. How many parameters will you have then? Explain if this structure would be better worse or the same as the previous one.

6. **Maximum likelihood**

Suppose you are given data $x_1, \ldots x_m$, where each x_i is a single real value (in other words, you have m instances and a single real-valued attribute). Suppose that the data is distributed *uniformly randomly* between $-w$ and w, You are trying to find the maximum likelihood estimate of w based on your data.

(a) Write the likelihood function $L(w)$.
Solution: $L(w) = \prod_{i=1}^{m} P_w(x_i)$ where:

$$P_w(x_i) = \begin{cases} \frac{1}{2w} & \text{if } x_i \in [-w, w] \\ 0 & \text{otherwise} \end{cases}$$

(b) What is the maximum likelihood estimate for w? Justify your answer based on the likelihood function

Solution: Note that if any x_i is outside $[-w, w]$, the whole likelihood will be 0. At the same time, we would like w to be as small as possible, to maximize $1/(w2)$. Hence, we should pick w to fit both the min and max data point that we are given:

$$w = \max(|\min_{1 \le i \le m}(x_i)|, |\max_{1 \le i \le m}(x_i)|)$$

(c) Suppose now that you have labelled data $\langle x_j, y_j \rangle$, where y_j is either $+1$ or 0. Recall that a generative classifier will model $P(y)$ and $P(x|y)$. Give an example of a data set for which the generative approach, using the model described above for each $P(x|y)$, will not work well.

Solution: Suppose that you have all the x for the positive class between 0 and some number M and all those for the negative class between $-M$ and 0. Then the ML estimate for x for both classes will end up with $w = M$, and as a result, you would always classify the example based on the prior over classes. Yet the examples are perfectly separable. The problem here stems from the bias of the classifier, which is required to be symmetric.

(d) Recall that a discriminative classifier will model directly $P(y|x)$. Could you solve the example you gave in the previous part using such a classifier? If so, explain what the classifier would look like.

Solution: Yes, you could simply pick a threshold and classify based on whether x is larger or smaller than the threshold. Note that one could also allow non-symmetric boundaries on the distributions, in which case the generative model would work as well.

7. **Short questions**

(a) Why are L_1 and L_2 regularization not used for support vector machines?

Solution: Because they use the max-margin optimization criterion with soft margins, which already provides regularization. L_1 and L_2 would work in feature space, but the provided regularizer works in the dual space, where we want to d the optimization.

(b) True or false: SVMs trained on the same set of data will always result in the same classifier, assuming the same C parameter value (justify in 1 sentence).

Solution: True, because we solve a convex optimization problem whose optimum, given a set of data and C, is unique.

(c) You want to use kernel-based logistic regression. In one sentence, when would you prefer L_1 regularization over L_2 regularization?

Solution: Kernel-based logistic regression works in the dual space, so will have a para,ether associated with every instance. L_1 regularization will drive some of these to 0, which means the distance is discard in effect from the dataset. Hence, we would use this approach when we want to "sparsity" the dataset.

(d) Suppose that you train a parametric classifier with training sets of size m. As $m \to \infty$, what do you expect will be the behavior of the training error? What would you expect for the behavior of the test error? You can draw a picture if it helps.

Solution: If the classifier is parametric, both the training and the test error should go down and then up, though the training error will go down more quickly. The error in the first part of the curve is dominated by the variance, in the latter part of the curve, by the bias.

(e) Suppose that you have a linear SVM binary classifier. Consider a point that is currently classified correctly, and is far away from the decision boundary. If you remove the point from the training set, and re-train the classifier, will the decision boundary change or stay the same? Explain your answer in one sentence.

Solution: The SVM decision boundary is defined by the support vectors. Due to regularization, if we only remove one point from the training set, and it was not a support vector, it is unlikely the boundary will change.

(f) Suppose that you have a logistic regression classifier. Consider a point that is currently classified correctly, and is far away from the decision boundary. If you remove the point from the training set, and re-train the classifier, will the decision boundary change or stay the same? Explain your answer in one sentence.

Solution: Since all points contribute to the decision boundary (which is defined by the

(g) True or false: $p(A) \geq p(A, B), \forall A, B$

Solution: True, since $p(A = a)$ is obtained by summing $p(a, b)$ for all values of $B = b$.

(h) You have a Bayes net with 100 nodes. Imagine the lexicographic ordering of the nodes, such that a node comes after all its parents in the ordering. You have a query conditioned on a variable that is almost at the end of this ordering. You want to use approximate inference to compute conditional probability queries for this variable. Should you use likelihood weighting / importance sampling or Gibbs sampling? Justify your answer in one sentence

Solution: Gibbs sampling is preferable, as you will generate samples in which this variable is always set correctly, and you can do this starting in its neighbourhood. Importance sampling will generate samples by going in the lexicographic ordering, and suddenly when you get to the conditioning variable, the weight is adjusted and will be small. Hence, one would expect Gibbs in this case to produce fewer, more "targeted" samples.

(i) Expectation maximization (EM) is designed to find a maximum likelihood setting of the parameters of a graphical model when some of the data is missing. Does the algorithm converge? If so, do you obtain a locally or globally optimal set of parameters?

Solution: EM converges, but to a locally optimal solution.

(j) Suppose you have data coming from an HMM, but you do not know the number of states. You try to run EM with increasing number of states in the HMM structure (1, 2, etc). Will this approach have a preference for a small or a large number of states? Justify your answer.

Solution: If we aim just to maximize likelihood, having more parameters allows higher likelihoods. Hence, this approach would favour a bigger number of states. We would need to use regularization of some sort (to encourage small models), or a validation set to prevent this.

8. **Naive Bayes**

Suppose that you are trying to solve a binary classification problem, and your data set has 4 attributes. Each attribute can take 3 possible values.

(a) If you modeled the full joint distribution of the attributes and the class label, how many parameters would you need?

Solution: We would have $P(X_1, X_2, X_3, X_4, Y)$, which is a table with $3^4 * 2$ entries, of which $3^4 * 2 - 1$ are independent parameters.

(b) If instead you use a Naive Bayes classifier, how many parameters will you have to fit?

Solution: We have one parameter for the class. Then, for each X_i, we have $3 - 1 = 2$ parameters for each class, so 4 parameters associated with each attribute node. In total, $4 * 4 + 1 = 17$ parameters.

9. **Maximum likelihood and gradient descent**

Suppose that you want to train a regressor to maximize the likelihood of the training data, under the usual Gaussian noise assumption.

(a) Would minimizing the mean-squared error function make sense in this case? If yes, justify your answer. If not, give an alternative error measure that should be minimized

Solution: Yes (See lecture on the probabilistic interpretation of mean-squared error, and homework 1).

(b) Suppose that your hypothesis has the form $h_{\mathbf{a},\mathbf{b}}(\mathbf{x}) = \sum_{j=1}^{n} a_j \sin(b_j + x_j)$. Derive gradient descent update rules for parameters a_j, b_j, $j = 1 \ldots n$.

Solution: The objective function is:

$$J(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{m} (y_i - h_{\mathbf{a},\mathbf{b}}(\mathbf{x}_i))^2 = \sum_{i=1}^{m} (y_i - \sum_{j=1}^{n} a_j \sin(b_j + x_{ij}))^2$$

The weight update is:

$$a_j \leftarrow a_j - \alpha \frac{\partial J}{\partial a_j} = a_j + \alpha \sum_{i=1}^{m} \left[(y_i - \sum_{j=1}^{n} a_j \sin(b_j + x_{ij})) \sin(b_j + x_{ij}) \right]$$

(where we've folded the 2 from the derivative into the learning rate). Similarly:

$$b_j \leftarrow b_j - \alpha \frac{\partial J}{\partial b_j} = b_j + \alpha \sum_{i=1}^{m} \left[(y_i - \sum_{j=1}^{n} a_j \sin(b_j + x_{ij})) \cos(b_j + x_{ij}) \right]$$

(c) Will your gradient descent procedure converge to a unique solution? Justify your answer.

Solution: No - as we have periodic functions in the hypothesis, there will be in fact many values for the bs (spaced appropriately) which yield the same value of the objective, so it will have local optima.

(d) Consider the hypothesis class $h(\mathbf{x}) = \sum_{j=1}^{n} \log(e^{w_j x_j})$. Will gradient descent for finding the weights converge to a unique solution for this hypothesis? Justify your answer.

Solution: $h(\mathbf{x}) = \sum_{j=1}^{n} w_j x_j$ so this is just a linear hypothesis. Hence, the mean-squared error objective has one global optimum, and gradient descent with the appropriate conditions on the learning rate will reach it.

10. **Gibbs sampling**

Suppose you do Gibbs sampling in a Bayes net with no evidence. What does the Markov chain look like? (Give a qualitative description). What will be its stationary distribution?

Solution: Since we have no evidence, the states will consist of values assigned to all the variables in the network. The transitions are as follows: we pick uniformly at random a variable, then we re-sample it conditioned on the current values of the variables in its Markov Blanket. The stationary distribution is the joint probability over all the variables in the network (as represented in the Bayes net).

11. **Inference in Markov nets**

Let X be a node in a Markov network. and let y be an assignment of values to the nodes $Y = MarkovBlanket(X)$. Show that the ratio:

$$\frac{p(X = x'|y)}{p(X = x|y)}$$

can be computed efficiently, based only on local parameters. Explain how to find the most likely value of a variable given evidence about the parents based on this observation.

Solution:

$$\frac{p(X = x'|y)}{p(X = x|y)} = \frac{p(X = x', y)}{p(X = x, y)} = \frac{\prod_C \psi_C(X = x', y)}{\prod_C \psi_C(X = x', y)}$$

where C are the cliques over X and variables in Y and ψ are potentials. Hence, the computation is local (only involving the potential in which a node participates).

12. **Undirected models**

    Let X_i, $i = 1 \ldots n$ and Y be random variables. Draw a Markov network such that, for all $i \neq j$, $X_i \perp\!\!\!\perp X_j | Y$.

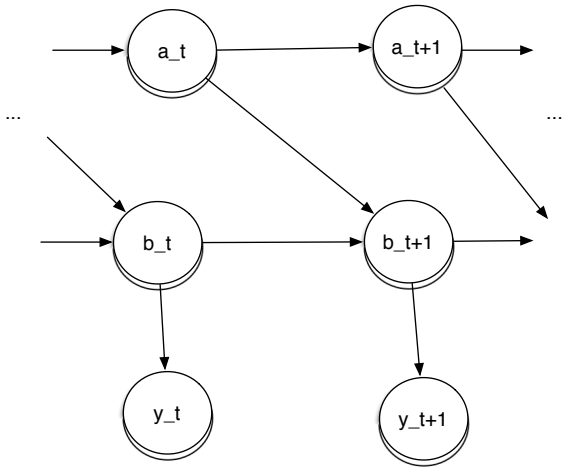
    Solution: the graph structure is a star with Y in the middle and all X_i connected only to Y.

13. **Undirected models** Suppose that someone asks you to draw a Markov network over variables A, B, C and D which obeys the following conditional independency properties: $A \perp\!\!\!\perp B | C$ and $A \not\perp\!\!\!\perp C | D$. Is this possible? If so, draw the network. If not, justify why.

    Solution: $A - D - C - B$. Removing C disconnects A and B. Removing D disconnects A and C.

14. **Hidden Markov Models**

    (a) Suppose that you have an HMM in which each state s is actually described by two state variables a and b. The value a_{t+1} depends probabilistically only on a_t. The value b_{t+1} depends probabilistically on both a_t and b_t. The observations y_t depend only on b_t, not on a_t. Draw the graphical model for representing a trajectory. Describe what parameters are needed to specify the HMM.



    The parameters are: $P(A_{t+1} = a' | A_t = a)$ (in the discrete case, if the domain of A has n values, this is an $n \times n$ matrix), $P(B_{t+1} = b' | A_t = a, B_t = b)$ (in the discrete case, this is

a $m \times (nm)$ matrix if B has m possible values), and $P(Y_t = y|B_t = b)$ (a $k \times m$ matrix, if there are k possible observations).

Notice that if we made a state variable s_t with components a_t and b_t, this would exactly be an HMM, but our model gives more structure (e.g. y_t only depends on part of the state).

(b) Write a forward algorithm for computing the probability of a state sequence given observations, for this HMM.

Solution: We want: $P(a_{t+1}, b_{t+1}|y_1, \ldots y_t)$ (aka the belief state, given the observation sequence so far). We assume $P(a_0, b_0$ is part of the model. We have:

$$P(a_{t+1}, b_{t+1}|y_1, \ldots y_t) = \frac{P(a_{t+1}, b_{t+1}, y_1, \ldots y_t)}{\sum_{a,b} P(a_{t+1} = a, b_{t+1} = b, y_1, \ldots y_t)}$$

The denominator is a normalization, so we focus on the numerator.

$$\begin{aligned} P(a_{t+1}, b_{t+1}, y_1, \ldots y_t &= \sum_{a_t} \sum_{b_t} P(a_{t+1}, b_{t+1}, a_t b_t, y_1, \ldots y_t) \\ &= \sum_{a_t} \sum_{b_t} P(a_{t+1}|a_t) P(b_{t+1}|a_t, b_t) P(y_t|b_t) P(a_t, b_t|y_1, y_{t-1}) \end{aligned}$$

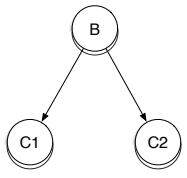
where we used the structure of the model, and the last term is the belief at the previous time step. Hence we have a belief update formula that can be applied forward (as we get more observations).

15. **Expectation maximization**

Suppose that somebody gave you a bag with two biased coins, having the probability of coming up heads of p_1 and p_2 respectively. You are supposed to figure out p_1 and p_2 by tossing the coins repeatedly. You will repeat the following experiment n times: pick a coin uniformly at random from the bag (i.e. each coin has probability $1/2$ of being picked) and toss it, recording the outcome (heads or tails). The coin is then returned to the bag. Assume that the individual experiments are independent.

(a) Set up a graphical model which captures this problem.

Solution:



(b) Suppose the two coins have different color: the p_1 coin is white, the other coin is yellow. Show the maximum likelihood estimators for the two parameters, p_1 and p_2.

Solution: p_1 is the frequency of observing heads when tossing the white coin, p_2 is the same for the yellow coin.

(c) Suppose now that the two coins look identical, so when you take them out of the bag you cannot tell them apart. Hence, the identity of the coin is always missing in your data. Write the expected log-likelihood of the data in this case.

Solution: Say that $B = 1$ means we're flipping $C1$ and $B = 2$ means flipping $C2$. The expected log-likelihood for a set of m instances x_i is:

$$\sum_{i=1}^{m} \log \left(\sum_b P(x_i, b) \right) = \sum_{i=1}^{m} \log \left(\sum_b (P(x_i|B=1)P(B=1) + P(x_i|B=2)(1 - P(B=1))) \right)$$

We can also write $P(x_i|B = 1) = p_1\delta_{x_i=H} + (1 - p_1)(1 - \delta_{x_i=H})$, and similarly for the $P(x_i|B = 2)$.

(d) Suppose you start with some guesses for the parameters, $\hat{p_1}$ and $\hat{p_2}$. Show the E-step of the soft EM algorithm.

Solution: In the E-step we fill in the missing values, so we will have:

$$w_{i1} = P(B = 1|x_i) = \frac{P(x_i|B=1)P(B=1)}{P(x_i|B=1)P(B=1) + P(x_i|B=2)P(B=2)}$$

and

$$w_{i2} = P(B = 2|x_i) = \frac{P(x_i|B=2)P(B=2)}{P(x_i|B=1)P(B=1) + P(x_i|B=2)P(B=2)}$$

If we let $\hat{p}_0 = P(B = 1)$, note $P(B = 2) = 1 - \hat{p}_0$ and we have $P(x_i|B = 1) = \hat{p}_1\delta_{x_i=H} + (1 - \hat{p}_1)(1 - \delta_{x_i=H})$, and similarly for $P(x_i|B = 2)$.

(e) Show the M-step of the EM algorithm.

Solution:

$$\begin{aligned} \hat{p}_0 &= \frac{1}{m} \sum_{i=1}^{m} w_{i1} \\ \hat{p}_1 &= \frac{\sum_{i=1}^{m} \delta_{x_i=H} w_{i1}}{\sum_{i=1}^{m} w_{i1}} \\ \hat{p}_2 &= \frac{\sum_{i=1}^{m} \delta_{x_i=H} w_{i2}}{\sum_{i=1}^{m} w_{i2}} \end{aligned}$$

(feel free to simplify this further and see what you get)