TDA 231 Machine Learning 2015: Final Exam

Instructor: Devdatt Dubhashi

Due:10 AM, Room 6453, March 19, 2015

1. (10 points) Observations $y_1, \cdots, y_n$ are i.i.d. noisy measurements of an underlying sensor variable x with $P(y_i \mid x) \sim \mathcal{N}(x, \sigma^2)$ with known σ^2.
   (a) Write the likelihood function.
   (b) Select a prior distribution on x which is conjugate to the likelihood.
   (c) Write the posterior distribution explicitly giving the formulas for the parameters (Hint: use the conjugacy property).

2. (10 points) A local supermarket specializing in breakfast cereals decides to analyse the buying patterns of its customers. They make a small survey asking six randomly chosen people their age (older or younger than 60) and which of the breakfast cereals (Cornflakes, Frosties, Sugar Puffs, Branflakes) they like. Each respondent provides a vector with 0s and 1s corresponding to whether they like or dislike the cereal. Thus a respondent with (1101) would like Cornflakes, Frosties and Branflakes but not Sugar Puffs. The data obtained was:

   >**60** (1000), (1001), (1111), (0001).

   <**60** (0110), (1110).

   A new customer comes into the supermarket and says she only like Frosties and Sugar Puffs. Using naive Bayes trained with maximum likelihood, what is the probability she is younger than 60?
   (a) Write the likelihood and list the parameters you need to estimate.
   (b) Write the equations for the MLE estimates of the parameters.
   (c) Compute the probability the new customer is younger than 60.

3. (15 points) A sequence of dice throws is generated by the following process involving a coin and two die, A and B: first we toss the coin. If it comes up `heads`, we select dice A else if `tails` we select dice B. Now the selected dice is tossed and the outcome observed. Only the final outcome – namely a face between 1 and 6 – is observed i.e. we do not observe the result of the coin toss. The problem is to estimate the bias of the coin θ and the loading of the die given by the probability distributions on the six faces θ^A and θ^B respectively, based on the observed data. Let $\mathbf{w} := w_1, w_2, , \cdots, w_N$ be the observed dice throws, and let $\mathbf{z} := z_1, z_2, \cdots, z_N$ denote the unobserved coin tosses with each $z_i \in \{A, B\}$.
   (a) Write the likelihood function $P(\mathbf{z} \mid \theta)$ and choose a suitable prior that is conjugate to it.
   (b) Write the likelihood function $P(\mathbf{w} \mid \mathbf{z}, \theta^A, \theta^B)$, and choose (independent) priors for the die conjugate to the likelihood.
   (c) Write the $P(\mathbf{w} \mid \theta, \theta^A, \theta^B)$ and the posterior distribution $P(\theta, \theta^A, \theta^B \mid \mathbf{w})$ using Bayes's rule and comment on the difficulty of computing it.

   Design a MCMC algorithm to sample from the posterior $P(\mathbf{z} \mid \mathbf{w})$ following the steps given below.

(d) Write down the transition probability for a Gibbs sampler on the Markov chain whose states are given by $\mathbf{z}$ so that the stationary distribution is the posterior distribution i.e. write down the transition probabiltiies corresponding to the component z_i i.e. the conditional probability distribution $P(z_i \mid \mathbf{z}_{-i}, \mathbf{w})$. You may need to use the notation for counts: $n_A := \sum_i [z_i = A]$, and $n_{A,f}$ for the number of i where $z_i = A$ and $w_i = f$, where $f \in \{1 \cdots, 6\}$ etc. (Hint: Recall LDA exercise!)

(e) State what are the distributions $P(\theta, \mid \mathbf{z}, \mathbf{w})$, $P(\theta^A \mid \mathbf{z}, \mathbf{w})$ and $P(\theta^B \mid \mathbf{z}, \mathbf{w})$. Using this, state how, given a sample $\tilde{\mathbf{z}}$ from the posterior, you can get samples from an approximation of the posterior distribution $P(\theta, \mid \mathbf{w})$, $P(\theta^A \mid \mathbf{w})$ and $P(\theta^B \mid \mathbf{w})$.

4. (10 points) Same setup as previous problem: A sequence of dice throws is generated by the following process involving a coin and two die, A and B: first we toss the coin. If it comes up `heads`, we select dice A else if `tails` we select dice B. Now the selected dice is tossed and the outcome observed. Only the final outcome namely a face between 1 and 6 is observed i.e. we do not observe the result of the coin toss. The problem is to estimate the bias of the coin θ and the loading of the die given by the probability distributions on the six faces θ^A and θ^B respectively, based on the observed data. Let $\mathbf{w} := w_1, w_2, , \cdots, w_N$ be the observed dice throws, and let $\mathbf{z} := z_1, z_2, \cdots, z_N$ denote the unobserved coin tosses with each $z_i \in \{A, B\}$.

(a) Write the form of the joint probability distribution $P(\mathbf{z}, \mathbf{w} \mid \theta, \theta^A, \theta^B)$.

(b) Write the E step update: assuming we have good guesses of the parameters $\theta, \theta^A, \theta^B$, update the beliefs on the latent variable $\mathbf{z}$. Give both a hard and a soft version of the update.

(c) Write the M step updates: assuming we know the latent variable $\mathbf{z}$, what are the MLE estimates of the parameters? Give the estimates correpsonding to both the hard and soft versions of the E step.

(d) How would you initialize? Will the algorithm always return the same answer?

(e) Compare the pros and cons of this approach versus the MCMC approach in the previous problem.

5. (15 points) Consider first the following binary training data:

+1 : $(3,1), (3,-1), (6,1), (6,-1))$

-1 : $(1,0), (0,1), (0,-1), (-1,0)$.

(a) Plot the points and draw the optimal separating line by inspection.

(b) Write the primal and dual SVM problems *correpsonding to this instance* (not the general formulation!). Give the optimal solutions to the primal and dual and indicate the support vectors. Write the separating hyperplane in terms of the support vectors.

Now consider the following data:

+1 : $(2,2), (2,-2), , (-2,-2), (-2,2)$

-1 : $(1,1), (1,-1), (-1,-1), (-1,1)$.

(c) Plot the points. Are the classes linearly separable?

(d) Consider the mapping

$$\Phi(x_1, x_2) := \begin{cases} (4 - x_2 + |x_1 - x_2|, 4 - x_2 + |x_1 - x_2|) & \text{if} \sqrt{x_1^2 + x_2^2} > 2 \\ (x_1, x_2) & \text{otherwise} \end{cases}$$

Plot the transformed points $\Phi(\mathbf{x}_i)$ – are they linearly separable?

(e) Write down the Gram matrix of the corresponding kernel and write the dual SVM programme for this instance.

(f) Give the optimal solution to the dual.

(g) Write down the optimal separating hyperplane $\mathbf{w}^T\Phi(\mathbf{x}) = 0$ explicitly for this problem instance using (f) and use it to classify the point $(4, 5)$ in the original (non–transformed space).

(h) Consider the mapping

$$\Phi'(x_1, x_2) := (x_1, x_2, (x_1^2 + x_2^2 - 5)/3)$$

(i) Give the Gram matrix for Φ' and formulate the corresponding dual.

(j) Write down the optimal separating hyperplane $\mathbf{w}^T\Phi'(\mathbf{x}) = 0$ explicitly for this problem instance and use it to classify the point $(4, 5)$ in the original (non–transformed space)