TDA 231 Machine Learning 2016: Final Exam

Instructor: Devdatt Dubhashi

Due: 4 PM, Room 6472, March 16, 2016

1. (10 points) A sequence of points $(x_1, y_1), \cdots, (x_N, y_N)$ is described by the following model:

$$y_i = mx_i + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

   and each data point is independent of the others.

   (a) Write the likelihood function $P(y_1, \cdots y_n \mid x_1, \cdots, x_n, m, \sigma^2)$.

   (b) Compute the MLE estimate of m.

   (c) Select a prior distribution on σ^2 which is conjugate to the likelihood.

   (d) Write the posterior distribution explicitly giving the formulas for the parameters (use conjugacy!).

2. (10 points) Consider a 3–class Naive Bayes classifier with one binary and one Gaussian feature:

$$y \sim \mathrm{Cat}(\pi), \quad x_1 \mid y = c \sim \mathrm{Ber}(\theta_c), \quad x_2 \mid y = c \sim \mathcal{N}(\mu_c, \sigma_c^2).$$

   (Recall definitions of categorical, Bernoulli and Gaussian variables!)

   (a) Write the joint distribtuion.

   Suppose the parameters are:

$$\pi = (0.5, 0.25, 0.25), \theta = (0.5, 0.75, 0.5), \mu = (-1, 0, 1), \sigma^2 = (1, 1, 1).$$

   (b) Compute $P(y \mid x_1 = 0, x_2 = 0)$ (the result should be a vector whose entries sum to 1). Show your reasoning

   (c) Compute $P(y \mid x_1 = 0)$. Show your reasoning.

   (c) Compute $P(y \mid x_2 = 0)$. Show your reasoning.

3. (10 points) A sequence of points $(x_1, y_1), \cdots, (x_N, y_N)$ is produced by the following generative model. There are L straight lines $y = m_\ell x, \ell = 1 \cdots L$. To generate a point:

   1. Pick $\ell \in \{1 \cdots L\}$ uniformly at random.
   2. Generate $x \sim \mathcal{N}(\mu_\ell, 100)$.
   3. Generate $y = m_\ell x + \mathcal{N}(0, \sigma^2)$.

   The aim of this probem is to infer the underlying model from the data.

   (a) Draw the probabilistic graphical model representing this process. Adopt a Bayesian approach allowing for priors on parameters, and use plate notation.

   (b) Write the joint distribution function represented by it.

   (c) Describe a Markov chain that explores the parameter space - what are the states of this Markov chain?

   (d) Outline Metropolis–Hastings transitions on this Markov chain to sample from the posterior.

4. (10 points) Similar setup as previous problem. Points $(x_1, y_i) i = 1 \cdots N$ are generated as follows:

   1. Choose line ℓ with probability p_ℓ, for $\ell \in \{1 \cdots L\}$.
   2. Generate $x \sim \mathcal{N}(0, 100)$ and $y \sim m_\ell(x) + \mathcal{N}(0, \sigma^2)$

   In this problem you will develop a EM style algorithm to estimate the hidden.

   (a) Write the E step assuming all parameters are known: compute the *residuals* $\Delta_\ell := (y - m_\ell x)^2$ for each point with respect to line ℓ and use a *softmax* assignment based on these residuals.

   (b) Write the M step assuming the assignment of points to lines is known. What are the MLE estimates of the parameters?

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

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

5. (10 points) Consider a neural network with a single hidden layer of logistic units being used for a multi–class classification problem:

$$\mathbf{h} = \sigma(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}), \quad \hat{\mathbf{y}} = \mathrm{softmax}(\mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)}).$$

   and trained using the cross–entropy error:

$$C(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_i y_i \log \hat{y}_i.$$

   (a) If the input is D dimensional, the number of classes is k and the number of hidden units is H, what is the total number of parameters in the model?

   (b) Write down the gradients of the error with respect to the parameters in the first layer, i.e. the layer closest to the input. Assume the output target $\mathbf{y}$ is a one–hot representation. You may find the following useful: $\frac{\partial C}{\partial \mathbf{z}} = \mathbf{y} - \hat{\mathbf{y}}$, where $\mathbf{z} = \mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)}$.

6. (10 points) Consider first the following binary training data:

   $+\mathbf{1}$: $(4, 4), (4, 0), (2, 2), (0, 0)$

   -1 : $(2, 0), (0, 2)$.

   This is the same as in your problem set *except* that the point $(0, 0)$ was in class -1. In that case, you computed the optimal maximum margin separator to be the line $x_1 + x_2 - 3 = 0$.

   (a) It is visually clear that the data set is not linearly separable. How would you prove this? That is, show that *no* line can separate the two classes.

   (b) Write the primal and dual *soft margin* SVM formulations *correpsonding to this instance*. Do not use the general formulation, do not use summation signs.

   (c) For $C = 10$, write down values of slack variables in the primal corresponding to a feasible solution using the line $x_1 + x_2 - 3 = 0$. What is the primal objective value? Is this the optimal solution?

   (d) For $C = 1$, find the optimal solution, give the objective function value, plot the seprating line and indicate the support vectors. Write the separating hyperplane in terms of the support vectors.

   (e) True of false? "If you apply a soft margin SVM to a linearly separable data set you recover the hard margin separator". Justify briefly.