# COL 774: Machine Learning
# Major Examination

Saturday May 6, 2017

**Notes:**

- **Time: 10:30 pm to 12:35 pm. Total Questions: 9. Maximum Points: 40. No. of Pages: 3.**
- **Each question carries 5 points. Attempt any 8 questions.**
- **Clearly mark the question you do not want to be graded.**
- **Start each answer on a new page. This exam is closed books & notes.**
- **Some questions may be harder than others. Use your time wisely.**
- **You need to justify all your answers. Answers without justification may not get full/any points.**

1. Consider a 3-d volume of size $w \times h \times d$ in a CNN layer ($w$:width,$h$:height, d:depth). Consider applying $r$ convolutional kernels of size $k \times k$ to this volume. Let $s$ denote the length of the stride i.e., the number of pixels by which the kernel window is moved (either horizontally or vertically) before next application. Answer the following.

   (a) What is the number of parameters to required implement the above convolutional layer. Do not forget the bias term(s). Briefly justify your answer.

   (b) What is the size of the resulting volume? You can assume that each of $w, h, k$ is divisible by $s$. Briefly justify your answer.

   (c) A convolutional layer such as above is typically followed by a pooling layer. Explain why.

2. In class, we defined the MAP estimator as $\theta_{MAP} = \arg\max_\theta P(\theta|D)$ i.e., the parameter vector $\theta$ whose probability is maximized given the data. Given a new instance $x$, the prediction probability is given by $P(y|x, \theta_{MAP})$, i.e. probability of label $y$ given feature vector $x$ and $\theta_{MAP}$ as the set of parameters. Now, consider the quantity, $P(y|x, D)$ which directly tries to estimate the probability of $y$ given $x$ conditioned on the data.

   (a) Show that $P(y|x, D)$ can be written as $\int P(y|x, \theta)P(\theta|D)d\theta$. You need to justify every step in your derivation. This is called Bayesian Averaging.

   (b) Consider a binary classification problem, i.e. $y \in \{0, 1\}$. Let $y_{MAP} = \arg\max_y P(y|x, \theta_{MAP})$, and $y_{AVG} = \arg\max_y P(y|x, D)$. Note that in our setting, for every parameter vector $\theta$, there is a corresponding (unique) hypothesis $h_\theta$. Consider a finite-sized hypothesis class $\mathcal{H} = \{h_1, h_2, h_3\}$ such that $P(h_1|D) = 0.3$, $P(h_2|D) = 0.4$ and $P(h_3|D) = 0.3$. Further, given an input $x$, let $P(y = 1|x, h_1) = 0.4$, $P(y = 1|x, h_2) = 0.6$ and $P(y = 1|x, h_3) = 0.4$. Compute $y_{MAP}$ and $y_{AVG}$. Are they equal? Explain. Note that since the hypothesis class is finite, in Bayesian averaging, you will have to perform a sum instead of an integral.

3. Recall that softmax regression is an extension of logistic regression for the case of a multi-class classification problem. Here is a quick recap. Assume that the target variable $y$ takes the value in the set $\{1, 2, \cdots, r\}$ (i.e., consisting of $r$ different labels). Let the input vector be given as $x \in \mathcal{R}^n$. The problem is now characterized by a set of parameter vectors $\Theta = \{\theta_{(k)}\}_{k=1}^{r}$ where each $\theta_{(k)} \in \mathcal{R}^{n+1}$. The probability distribution of $(y = k|x)$ is defined as:

$$P(y = k|x; \Theta) = \frac{1}{Z} * e^{(\theta_{(k)})^T x} \tag{1}$$

Here, $Z = \sum_k e^{(\theta_{(k)})^T x}$ is the normalization constant. For the problem below, assume that $r = 3$.

(a) How would you compute the decision boundary enforced by the above classifier? You should describe your answer in mathematical terms.

(b) For this part only, assume that $x \in \mathcal{R}^2$ (and $r = 3$ as earlier). Pictorially depict the decision surface learnt by a softmax classifier. You should justify your drawing.

4. Recall the paper that we studied in class on use of PCA for building oil vulnerability index. Assume that each country in the dataset is described by a set of seven indicators $\{x_1, x_2, \cdots, x_7\}$. Let $\{f_1, f_2, \cdots, f_7\}$ denote the 7 (normalized) principal components and $\lambda_1, \lambda_2, \cdots \lambda_7$ be the corresponding eigenvalues (in decreasing order).

(a) Derive the expression for computing the oil vulnerability index $OVI_k$ for a country $k$.

(b) How would you compute the contribution of each of the features $x_i$ $(1 \leq i \leq 7)$ in $OVI_k$?

5. Consider the set of axis-perpendicular hypercubes in $\mathcal{R}^n$ where each face of the hypercube is perpendicular to one of the axis. Given the dimension $x_j$ $(1 \leq j \leq n)$, the two faces of the cube lying perpendicular to the dimension $x_j$ are given by the equations of the form $x_j = \alpha_j$ and $x_j = \alpha'_j$, where $\alpha_j$ and $\alpha'_j$ are some scalars. For example, in $\mathcal{R}^2$, this is the set of rectangles each of whose sides are parallel to $x_1$ or $x_2$ axis (and perpendicular to the other). In $\mathcal{R}^3$, this is the set of cubes each of whose face is parallel to $x_2$-$x_3$, $x_3$-$x_1$ or $x_1$-$x_2$ plane (and perpendicular to $x_1, x_2, x_3$ axis, respectively). Given an axis perpendicular hypercube in $\mathcal{R}^n$, we define a hypothesis $h_C$ as $h_C(x) = \mathbb{1}\{x \in C\}$, i.e., $h_C(x)$ is 1 if $x$ lies inside the corresponding hypercube $C$, and 0 otherwise. Consider the class $\mathcal{H}$ consisting of hypothesis $h_C$'s as defined above. Show that VC-dimension of $\mathcal{H}$ is $2n$.

6. Consider training a linear SVM with the training set $\{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$. Assume that the training data is linearly separable. Further, assume that data is noise free and we use the hard-margin SVM model without any slack variables (i.e. no penalty terms). Let $|SV|$ be the number of support vectors obtained when training on the entire training set. Recall we say that $x^{(i)}$ is a support vector if and only if the correponding Langrange multiplier $\alpha_i > 0$. Let $\epsilon$ denote the $m$-fold cross-validation error (also called the leave-one-out error) of our SVM (i.e. the error obtained by training on a subset of $m-1$ points, testing on the remaining one and then averaging over all such $m-1$ sized subsets). Prove that $\epsilon \leq \frac{|SV|}{m}$. You need to justify every key step of your proof. Hint: Think about for which of the cross-validation folds the boundary learned on the $n-1$ points can be (potentially) different from the one learned for the entire training data (and why?). You may want to do this in the dual space.

7. Consider learning a perceptron with sigmoid as the activitation unit. Let the input to the network be specified as $x \in \mathcal{R}^n$. Let $\theta$ denote the weight vector. In the standard setting, $\theta \in \mathcal{R}^{n+1}$, i.e., we have a set of $n+1$ parameters (including the bias term). Recall that the error metric is given as $J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$ where $h_\theta(x)$ is as defined in class. Now, let us change the problem setting where the weights are tied to each other in groups of $r$ where $n$ is divisible by $r$. Let $k = \frac{n}{r}$. In other words, we enforce that $\theta_{l*r+1} = \theta_{l*r+2} = \cdots = \theta_{l*r+r}$, $\forall l, 0 \leq l \leq k-1$. Further note that we still have a separate $\theta_0$ term for bias. Therefore, we can alternately formulate the problem using a parameter vector $\theta'$ where $\theta' \in \mathcal{R}^{k+1}$.

| $x^{(i)}$ | $P(z^{(i)}=1)$ | $P(z^{(i)}=2)$ |
|---|---|---|
| 5 | 0.2 | 0.8 |
| 15 | 0.2 | 0.8 |
| 25 | 0.8 | 0.2 |
| 30 | 0.9 | 0.1 |
| 40 | 0.9 | 0.1 |

Table 1: E-step Probabilities

(a) Formulate the percpetron objective using the new set of parameters $\theta'$ as described above. Derive the gradient update rule for $\theta'$ in this new setting.

(b) Consider adding a regularizer term of the form $\lambda * \theta'^T\theta'$ to the above objective. How does the gradient change in this case?

8. Consider a learning problem with $n$ features satisfying the Naïve Bayes assumption i.e. $P(x|y) = \prod_{j=1}^{n} P(x_j|y)$. Let the target variable be $y \in \{0,1\}$ with $P(y=1) = \phi$. Let each feature $x_j$ be continuous valued with a Gaussian distribution conditioned on the class variable $y$. In particular, for the class $y=0$, $P(x_j|y=0) \sim \mathcal{N}(\mu_{j|0}, \sigma^2_{j|0})$ where $\mu_{j|0}$ is the mean of the distribution and $\sigma^2_{j|0}$ is its variance. Note that each variable $x_j$ has its own mean $\mu_{j|0}$ and variance $\sigma^2_{j|0}$. Similarly, for class $y=1$, we have $P(x_j|y=1) \sim \mathcal{N}(\mu_{j|1}, \sigma^2_{j|1})$. Above model is called Gaussian Naïve Bayes model.

(a) Show that Gaussian Naïve Bayes is a special case of GDA. Clearly describe the relationship between the two sets of parameters.

(b) Describe the kind of boundary learned by the Gaussian Naive Bayes model based on the relationship between the parameters $\{\mu_{j|0}, \sigma^2_{j|0}\}$ and $\{\mu_{j|1}, \sigma^2_{j|1}\}$.

Recall: GDA parameters are given as $\Theta = (\phi, \mu_0, \mu_1, \Sigma_0, \Sigma_1)$ where the symbols are as defined in class. In GDA, we first sample $y \sim Bernoulli(\phi)$. Then, we sample $(x|y=k) \sim \mathcal{N}(u_k, \Sigma_k)$ $(k \in \{0,1\})$. Also, if $x \sim \mathcal{N}(\mu, \Sigma)$, then $P(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp(\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$

9. Consider learning a GMM model with $k=2$ components. Further, assume that each input point $x \in \mathcal{R}$. Let there be 5 data points given as $\{5, 15, 25, 30, 40\}$. We will make the assumption that both the mixtures have the same co-variance matrix. Consider running EM algorithm over these points to estimate the paramters of the model. Assume that after a certain $E$ step run of the algorithm, the probabilities are given by Table 1.

(a) Compute the parameters learned in the next M step.

(b) Compute the new probabilities using the parameters obtained in the M step above.