

# END SEMESTER EXAM (CS-419)

25-Nov-2013

## 1 True/False Questions

**Note:** Clearly indicate whether the following statements are **True** or **False**. Marks will be awarded *only* if appropriate and compact justification (1-2 sentences) is provided.

1. *Sumati* posed the problem of spam-email filtering as a binary classification problem and decided to employ Support Vector Machines (SVMs) for this. Since the ideal kernel induced gram-matrix to be employed in the SVM is not known, he wishes to *learn* it using the training data provided for the classification problem at hand (and ofcourse domain knowledge). The problem of learning the gram-matrix for spam-email filtering is an instance of Supervised Learning.
2. *Sanjeevaka* observed that his SVM model with a polynomial kernel of degree 3 is over-fitting the training data. He should consider using a linear kernel.
3. The Bayesian Model Averaging (BMA) method fits/explains the training data better than the Maximum A-Posteriori (MAP) estimate based method.
4. While Naive Bayes classifier and Hidden Markov Model (HMM) are examples of generative models, Logistic and linear regression are examples of discriminative models.
5. The decision boundary with Quadratic Discriminant Analysis (QDA) and tied/shared covariance matrices<sup>1</sup> is linear.
6. *Karataka* always employs linear models, while *Damanaka* always employs quadratic models<sup>2</sup>. *Pingalaka*, being the king, decided to evaluate both over several datasets using the method of maximum marginal likelihood. It is more likely that *Damanaka* wins overall<sup>3</sup>.

---

<sup>1</sup>Covariance matrix of each class is the same (but unknown).

<sup>2</sup>Quadratic models include linear models.

<sup>3</sup>Assume that *Karataka*'s trained model is comparable to that of *Damanaka* in terms of likelihood of training data.

7. Ridge-regression is an example of a support vector method.
8. The Maximum Likelihood Estimate (MLE) for Gaussian variance, with known mean, is an unbiased estimator.
9. *Dashabuddhi*, *Shatabuddhi* and *Sahasrabuddhi* decided to employ cross-validation for tuning a hyper-parameter,  $C$ , in their model. *Dashabuddhi* used ten values of hyper-parameters, while *Shatabuddhi* and *Sahasrabuddhi* used 100 and 1000 values respectively<sup>4</sup>.
  - (a) *Kachadruma* evaluated their tuned models over the training data. It is most likely that *Sahasrabuddhi* will win and will be followed by *Shatabuddhi*.
  - (b) *Shibi* evaluated their tuned models over an unseen test data. It is most likely that *Sahasrabuddhi* will win and will be followed by *Shatabuddhi*.
10. For exponential family of distributions,
  - (a) the logarithm of the partition function is also the moment generating function.
  - (b) the partition function is convex.
11. The k-means algorithm is guaranteed to converge to a local maximum of the likelihood function.
12. If  $k$  is a (valid) kernel, then all entries in a gram-matrix induced by it will be non-negative.
13. Given a set of vectors  $\mathcal{X}$ , let  $\mathcal{I}$  and  $\mathcal{K}$  denote the set of all inner-products and kernels defined over  $\mathcal{X}$  respectively. Then<sup>5</sup>,  $\mathcal{I} \subset \mathcal{K}$ .
14. Let  $k$  be a kernel defined over  $\mathbb{R}^n$ . Let  $\mathcal{H}$  denote a Hilbert space where the given kernel  $k$  evaluates the inner-product. Then, the dimensionality of  $\mathcal{H}$  is greater than or equal to  $n$ .
15. Adaboost is a greedy iterative algorithm for minimizing empirical risk computed using square loss.

[15x1Mark=15Marks]

---

<sup>4</sup> Assume that the 10 ten values used by *Dashabuddhi* are included in the 100 values used by *Shatabuddhi* and in-turn included in the 1000 values used by *Sahasrabuddhi*.

<sup>5</sup>  $A \subset B$  denotes that  $A$  is a strict subset of  $B$ .

## 2 Analytical Questions

**Note:** Please write clear and legible answers. Marks *may not* be awarded if the hand-writing is un-readable. Minimize the number of english sentences and maximize mathematical sentences.

1. Provide a detailed description of the EM algorithm for MLE of a HMM. More specifically,
  - (a) Motivate and re-derive the EM algorithm as done in lecture.

[5Marks]
  - (b) Summarize the key steps (Pseudo-code).

[2Marks]
  - (c) Provide detailed description of a polynomial time algorithm<sup>6</sup> for computing the E-step.

[5Marks]
  - (d) Assuming the emission distributions are Gaussian, provide formulae for the M-step.

[2Marks]
  - (e) Provide polynomial time algorithm for computing the likelihood of training data once the model is trained.

[1Mark]
2. It is proposed to evaluate the popularity of Indian celebrities by measuring the popularity of their YouTube channels. One way to measure popularity of a channel is by simply modeling the distribution of: the sum of number of “likes” and “dislikes” in its videos. Higher the mean of these sums, higher is the popularity of the celebrity. A naive way to do this is to model each celebrity/channel by a Gaussian distribution or by a Gaussian likelihood and a suitable conjugate prior. However, since all the channels belong to a particular community, Indians, there will be latent factors that connect/tie all of them<sup>7</sup>. Such factors should be taken into account especially if the number of videos in each channel are less. As mentioned in lectures, one way to connect/tie these multiple models is by using a common prior<sup>8</sup>.

In summary, here is the description of the model: for the  $j^{th}$  celebrity, the model is Gaussian with mean  $\Theta_j$  and variance  $\sigma^2$ . The mean  $\Theta_j$  is what finally has to be estimated to decide who is popular. Assume that

---

<sup>6</sup>You need not follow the terminology, notation of the textbook/lectures nor re-produce the exact algorithm in the textbook/lectures. Any polynomial time algorithm would do.

<sup>7</sup>For eg. perhaps all Indian channels get less viewership than say US channels etc.

<sup>8</sup>Multi-task learning via Hierarchical Bayes method.

the variance  $\sigma^2$  is known. Now, the key modeling step is: we assume each  $\Theta_j$  to come from a common Gaussian prior with mean  $\mu$  and variance  $\tau^2$ . Along with  $\Theta_j$ , the hyper-parameters  $\mu$  and  $\tau^2$  are to be estimated. Assume that we employ Empirical Bayes<sup>9</sup> for parameter estimation. Derive the final simplified formula for the Empirical Bayes estimate of  $\Theta_j$  in terms of training data and  $\sigma^2$ .

[10Marks]

### 3 Numerical Questions

**Note:** Here you have to actually use the numerical data given and provide answers in terms of numbers. Your final answer should *not* be an analytical expression. Needless to say, you *should* show your working.

1. Recall the pmf of Poisson distribution  $p(x|\lambda) \equiv \frac{e^{-\lambda} \lambda^x}{x!}$ ,  $x = 0, 1, 2, \dots$ . Show that the Poisson distribution belongs to the exponential family by explicitly identifying the Partition function and the sufficient statistics. Now, assume that the following training data is given:  $\mathcal{D} = \{0, 0, 1, 0, 1, 2\}$ . Using your knowledge about the exponential family, or otherwise, compute  $p(3)$  with the following trained models:
  - (a) MLE model.
  - (b) MAP model<sup>10</sup>.
  - (c) BMA model.

Which of these values is higher? Is your finding intuitive?

[2+2+3+2+1=10Marks]

2. Consider the following binary classification<sup>11</sup> training data

$$\mathcal{D} = \left\{ \left( \begin{bmatrix} 1 \\ 0 \end{bmatrix}, +1 \right), \left( \begin{bmatrix} -1 \\ 0 \end{bmatrix}, +1 \right), \left( \begin{bmatrix} 0 \\ 1 \end{bmatrix}, -1 \right), \left( \begin{bmatrix} 0 \\ -1 \end{bmatrix}, -1 \right) \right\}$$

and the homogeneous quadratic kernel:  $k(x_1, x_2) \equiv (x_1^\top x_2)^2$ . Compute the discriminating hyperplane (i.e., the final prediction function<sup>12</sup>) obtained with a hard-margin SVM trained on this  $\mathcal{D}$  and this kernel  $k$ . *Hint:* Solve the SVM optimization problem using the geometry in it.

[2+2+3+2+1=10Marks]

<sup>9</sup>Recall that in Empirical Bayes, the (hyper) parameters are estimated using maximum (marginal) likelihood.

<sup>10</sup>Assume the appropriate conjugate prior.

<sup>11</sup>Labels are +1 or -1.

<sup>12</sup>Please simplify your final expression.