CSCI 567 Sp'21: Quiz 2 - Practice Discussion (28 April, 2021)

1 Multiple Choices Questions

Note that there are ONE or MORE correct answers to each question.

1. Which of the following is not a true statement about Lagrangian duality?

   (A) The solution of primal form and dual form is always equal.

   (B) They can be solved with convex optimization.

   (C) Duality lets us formulate optimality conditions for constrained optimization problems.

   (D) They can be optimized in the dual space.

   Solution: A

2. Which of the following statements about Expectation-Maximization (EM) algorithm are false?

   (A) Before running the EM algorithm, we need to choose the step size.

   (B) EM always converges to the global optimum of the likelihood.

   (C) In EM, the lower-bound for the log-likelihood function we maximize is always non-concave.

   (D) None of the above.

   Solution: A, B, C

3. Which of the following are true about generative modeling?

   (A) Naive Bayes and GMM are non-parametric generative models.

   (B) Generative models are generally non-parametric while discriminative models are generally parametric.

   (C) Generative models have more human-knowledge built into them than a corresponding discriminative model for the same problem.

   (D) Generative models can also generate data, while a discriminative model cannot.

   Solution: C, D

4. Which of the following statements about Gaussian Mixture Model (GMM) are true?

   (A) GMM is a non-parametric method that stores all the training samples.

   (B) GMM is a probabilistic model that can be used to explain how data is generated.

   (C) When learning a GMM, the labels of the samples are available.

   (D) After learning a GMM, one can infer a posterior distribution over the mixture components for a given test data point.

   Solution: B, D

5. Which of the following statements about Adaboost are true?

   (A) Adaboost does not need to reweight training data.

   (B) The Adaboost algorithm is resilient to overfitting.

   (C) AdaBoost may not output a linear classifier if the base classifiers are linear.

   (D) Non-linear classifiers cannot be used as base algorithms for Adaboost.

   Solution: B, C

2 Principal component analysis

Find the 1st and 2nd principal components of the dataset, whose centered covariance matrix is,

$$\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

Solution: Top principal components are eigenvectors of the centered covariance matrix.

- Solving for eigenvalues $Av = \lambda v$, so $|A - \lambda I| = 0$ since it has to be singular for $v \neq 0$. (Let A denote the centered covariance matrix).

$$\lambda = 2 + \sqrt{2}, 2, 2 - \sqrt{2}$$

(See for determinant of 3 x 3 matrix: https://www.mathsisfun.com/algebra/matrix-determinant.html)

- 1st and 2nd PC are eigenvectors corresponding to $\lambda_1 = 2 + \sqrt{2}, \lambda_2 = 2$. Substitute in $Av = \lambda v$ to get v, then normalize to unit norm (multiple possible answers).

$$v1 = [\frac{1}{2}, -\frac{\sqrt{2}}{2}, \frac{1}{2}]^\top, \quad v2 = [\frac{1}{\sqrt{2}}, 0, \frac{-1}{\sqrt{2}}]^\top$$

3 Naive Bayes Classifier

Suppose you are given the following set of data with three Boolean input variables a, b, and c, and a single Boolean output variable K. Assume we are using a naive Bayes classifier to predict the value of K from the values of the other variables

(a) According to the naive Bayes classifier, what is $P(K = 1|a = 1, b = 1, c = 0)$?

Solution: 1/2

$$\begin{aligned} P(K = 1|a = 1, b = 1, c = 0) &= \frac{P(K = 1, a = 1, b = 1, c = 0)}{P(a = 1, b = 1, c = 0)} \\ &= \frac{P(K = 1)P(a = 1|K = 1)P(b = 1|K = 1)P(c = 0|K = 1)}{P(a = 1, b = 1, c = 0, K = 1) + P(a = 1, b = 1, c = 0, K = 0)} \end{aligned}$$

where the naive bayes (independence assumption on a, b,c given K) is used.

a	b	c	K
1	0	1	1
1	1	1	1
0	1	1	0
1	1	0	0
1	0	1	0
0	0	0	1
0	0	0	1
0	0	1	0

(b) According to the naive Bayes classifier, what is $P(K = 0|a = 1, b = 1)$?

Solution: 2/3

$$\begin{aligned} P(K = 0|a = 1, b = 1) &= \frac{P(K = 0, a = 1, b = 1)}{P(a = 1, b = 1)} \\ &= \frac{P(K = 0)P(a = 1|K = 0)P(b = 1|K = 0)}{P(a = 1, b = 1, K = 0) + P(a = 1, b = 1, K = 1)} \end{aligned}$$

where the naive bayes (independence assumption on a, b given K) is used.

4 MLE and Expectation maximization

(a) An exponential distribution with parameter λ follows a distribution $p(x) = \lambda e^{-\lambda x}, \forall x \geq 0$. Given some i.i.d. data $\{x_i\}_{i=1}^N \sim \text{Exp}(\lambda)$, derive the maximum likelihood estimator $\hat{\lambda}$.

Solution: The log likelihood is

$$l(\lambda) = N \log \lambda - \lambda \sum_{i=1}^{N} x_i$$

Set the derivative to 0

$$\hat{\lambda} = \frac{N}{\sum_{i=1}^{N} x_i}$$

(b) Consider a mixture of two exponential distributions, with parameters λ_1, λ_2 respectively. Let $z \in \{0, 1\}$ be the hidden variable to indicate which exponential the observed data x is drawn from ($z = 0$ if $\text{Exp}(\lambda_1)$ is sampled, else $z = 1$ for $\text{Exp}(\lambda_2)$). Let the prior probability of $z = 1$ be denoted using parameter $\pi \in (0, 1)$, i.e., $p(z = 1) = \pi$.

Each example in the dataset $\{x_i\}_{i=1}^N$ is drawn i.i.d. from the mixture model. Write the E-step and M-step to estimate parameters $\lambda_1, \lambda_2, \pi$. Solve the M-step to obtain the update equations for the parameters, at each iteration of EM algorithm.

Solution: Posterior distribution is

$$\begin{aligned} P(z_i = 1|x_i; \lambda_1^t, \lambda_2^t, \pi^t) &= \frac{P(z_i = 1, x_i; \lambda_1^t, \lambda_2^t, \pi^t)}{P(z_i = 1, x_i; \lambda_1^t, \lambda_2^t, \pi^t) + P(z_i = 0, x_i; \lambda_1^t, \lambda_2^t, \pi^t)} \\ &= \frac{\pi^t \lambda_2^t e^{-\lambda_2^t x_i}}{\pi^t \lambda_2^t e^{-\lambda_2^t x_i} + (1 - \pi^t)\lambda_1^t e^{-\lambda_1^t x_i}} = \gamma_i^t \quad \text{(let us denote it using this symbol)} \end{aligned}$$

Then $P(z_i = 0|x_i; \lambda_1^t, \lambda_2^t, \pi^t) = 1 - \gamma_i^t$ (probability must sum to 1)

- E-step: Expectation of the complete log-likelihood with respect to the posterior distribution with parameters at iteration t (Let θ denote the set of parameters).

$$Q(\theta^t, \theta) = \sum_{i=1}^{N} \mathbb{E}_{z_i \sim p(.|x_i; \lambda_1^t, \lambda_2^t, \pi^t)} [\log P(x_i, z_i; \lambda_1, \lambda_2, \pi)]$$

$$\begin{aligned}
&= \sum_{i=1}^{N} \left[\gamma_i^t \log P(x_i, z_i = 1; \lambda_1, \lambda_2, \pi) + (1 - \gamma_i^t) \log P(x_i, z_i = 0; \lambda_1, \lambda_2, \pi) \right] \\
&= \sum_{i=1}^{N} \gamma_i^t \log(\pi \lambda_2 e^{-\lambda_2 x_i}) + \sum_{i=1}^{N} (1 - \gamma_i^t) \log \Big((1 - \pi) \lambda_1 e^{-\lambda_1 x_i} \Big) \\
&= \log \pi \sum_{i=1}^{N} \gamma_i^t + \log \lambda_2 \sum_{i=1}^{N} \gamma_i^t - \lambda_2 \sum_{i=1}^{N} \gamma_i^t x_i \\
&+ \log(1 - \pi) \sum_{i=1}^{N} (1 - \gamma_i^t) + \log \lambda_1 \sum_{i=1}^{N} (1 - \gamma_i^t) - \lambda_1 \sum_{i=1}^{N} (1 - \gamma_i^t) x_i
\end{aligned}$$

- M-step optimization problem: Maximize with respect to $(\lambda_1, \lambda_2, \pi)$ s.t. $0 < \pi < 1$.

$$\begin{aligned} &\arg\max_\theta Q(\theta^t, \theta) \\ &\text{s.t. } 0 < \pi < 1 \end{aligned}$$

- Solving the M-step: Setting derivative of Q with respect to parameters to 0 gives the following updates for the parameters (since after taking Lagrangian, with KKT complementary slackness the optimal dual variables are set to 0 for the constrained problem).

$$\pi = \frac{\sum_{i=1}^{N} \gamma_i^t}{N} \quad , \quad \lambda_2 = \frac{\sum_{i=1}^{N} \gamma_i^t}{\sum_{i=1}^{N} \gamma_i^t x_i} \quad , \quad \lambda_1 = \frac{\sum_{i=1}^{N} (1 - \gamma_i^t)}{\sum_{i=1}^{N} (1 - \gamma_i^t) x_i}$$

In the exam, the above argument with derivatives computation is enough to get parameter updates is enough. For ease of understanding, more detailed steps to the solution are outlined below. Rewriting the optimization as a minimization (not necessary, you can also write the Lagrangian from the maximization problem directly).

$$\begin{aligned} &\arg\min_\theta -Q(\theta^t, \theta) \\ &\text{s.t. } \pi - 1 \leq 0 \\ &\quad\quad -\pi \leq 0 \end{aligned}$$

You can write them as inequality constraints with $\leq$, then restrict the solution space later as $0 < \pi < 1$ (since this is a subset of that). With $\alpha, \beta \geq 0$ as the Lagrange multipliers, (and since objective and constraints are convex, we can use KKT as follows),

$$L = -Q(\theta^t, \theta) + \alpha(\pi - 1) + \beta(-\pi)$$

By complementary slackness $\alpha^*(\pi^* - 1) = 0$, and $\beta^*(-\pi^*) = 0$. Since $\pi^* \neq 0$ and $\pi^* \neq 1$, we must require $\alpha^* = \beta^* = 0$. With stationarity, we set derivative of L wrt primal variables to 0,

$$\frac{\partial L}{\partial \pi} = -\frac{\partial Q}{\partial \pi} + \alpha^* - \beta^* = 0$$

Therefore after taking derivatives, we get

$$\frac{\pi^*}{1 - \pi^*} = \frac{\sum_{i=1}^N \gamma_i^t}{\sum_{i=1}^N (1 - \gamma_i^t)} = \frac{\sum_{i=1}^N \gamma_i^t}{N - \sum_{i=1}^N \gamma_i^t}$$

Simplifying (adding the numerator to denominator on the left, and similarly for the right side of the equality), we get the update for $\pi^* = \frac{\sum_{i=1}^N \gamma_i^t}{N}$. Similarly derivative of Q with respect to λ_1, λ_2 will give the respective solutions for it.

5 Hidden Markov Models

Recall a hidden Markov model is parameterized by

- initial state distribution $P(Z_1 = s) = \pi_s$
- transition distribution $P(Z_{t+1} = s' \mid Z_t = s) = a_{s,s'}$
- emission distribution $P(X_t = o \mid Z_t = s) = b_{s,o}$

(a) Suppose we observe a sequence of outcomes $x_1, \ldots, x_T$ and would like to predict the next state Z_{T+1}, that is, we want to figure out for each possible state s,

$$P(Z_{T+1} = s \mid X_{1:T} = x_{1:T}).$$

Write down how one can compute this probability using the the forward message:

$$\alpha_s(T) = P(Z_T = s, X_{1:T} = x_{1:T}).$$

Solution:

$$\begin{aligned}
P(Z_{T+1} = s \mid X_{1:T} = x_{1:T}) &\propto P(Z_{T+1} = s, X_{1:T} = x_{1:T}) \\
&= \sum_{s'} P(Z_{T+1} = s, Z_T = s', X_{1:T} = x_{1:T}) && \text{(marginalizing)} \\
&= \sum_{s'} P(Z_T = s', X_{1:T} = x_{1:T}) P(Z_{T+1} = s \mid Z_T = s', X_{1:T} = x_{1:T}) \\
&= \sum_{s'} \alpha_{s'}(T) P(Z_{T+1} = s \mid Z_T = s') && \text{(Markov property)} \\
&= \sum_{s'} \alpha_{s'}(T) a_{s',s}
\end{aligned}$$

Therefore,

$$P(Z_{T+1} = s \mid X_{1:T} = x_{1:T}) = \frac{\sum_{s'} \alpha_{s'}(T) a_{s',s}}{\sum_{s''} \sum_{s'} \alpha_{s'}(T) a_{s',s''}}.$$

(b) More generally, suppose based on the same observation $x_1, \ldots, x_T$ we would like to predict the state at time $T + k$ for $k \geq 1$, that is, we want to figure out for each possible state s,

$$P(Z_{T+k} = s \mid X_{1:T} = x_{1:T}).$$

Write down how one can compute this probability by establishing a recursive form. In other words, express the above probability in terms of $P(Z_{T+k-1} = s' \mid X_{1:T} = x_{1:T})$ and the model parameters. Solution:

$$\begin{aligned}
P(Z_{T+k} = s \mid X_{1:T} = x_{1:T}) &= \sum_{s'} P(Z_{T+k} = s, Z_{T+k-1} = s' \mid X_{1:T} = x_{1:T}) \quad \text{(marginalizing)} \\
&= \sum_{s'} P(Z_{T+k-1} = s' \mid X_{1:T} = x_{1:T}) P(Z_{T+k} = s \mid Z_{T+k-1} = s', X_{1:T} = x_{1:T}) \\
&= \sum_{s'} P(Z_{T+k-1} = s' \mid X_{1:T} = x_{1:T}) a_{s',s} \qquad \text{(Markov property)}
\end{aligned}$$