CSCI 567, Spring 2023
Yan Liu

Homework #4

Released: Mar 28, 2023
Due: 11:59 pm, Apr 12, 2023 (AoE)
Total points: 65 points

Instructions

Notes on notation:

- Unless stated otherwise, scalars are denoted by small letter in normal font, vectors are denoted by small letters in bold font and matrices are denoted by capital letters in bold font.
- $\|.\|$ means L2-norm unless specified otherwise i.e. $\|.\| = \|.\|_2$

Problem 1 Hidden Markov Models (25 points)

Recall that a hidden Markov model (HMM) is parameterized as follows:

- 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 $P(X_t = o \mid Z_t = s) = b_{s,o}$

1.1 Suppose we observe a sequence of outcomes $x_1, \ldots, x_T$ and would like to predict the next state Z_{T+1} (represented as follows for a specific state s):

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

Show how this probability can be represented using the forward message

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

(15 points)

1.2 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 $k \geq 1$:

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

Write down how to compute this probability using a recursive form. That is, express the above probability in terms of $P(Z_{T+k-1} = s' \mid X_{1:T} = x_{1:T})$ and the model parameters. **(10 points)**

Problem 2 Principle Component Analysis (25 points)

In this problem, we use proof by induction to show that the M-th principle component corresponds to the M-th eigenvector of $\boldsymbol{X}^T\boldsymbol{X}$ sorted by the eigenvalue from largest to smallest. Here $\boldsymbol{X}$ is the centered data matrix and we denote the sorted eigenvalues as $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d$. In lecture, the result was proven for $M = 1$. Now suppose the result holds for a value M, and you are going to show that it holds for $M+1$. Note that the $M+1$ principle component corresponds to the solution of the following optimization problem:

$$\max_{\boldsymbol{v}} \quad \boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{v} \tag{1}$$

$$\text{s.t.} \quad \|\boldsymbol{v}\|_2 = 1 \tag{2}$$

$$\boldsymbol{v}^T\boldsymbol{v}_i = 0, i = 1, ..., M \tag{3}$$

where $\boldsymbol{v}_i$ is the i-th principle component. Write down the Lagrangian of the optimization problem above, and show that the solution $\boldsymbol{v}_{M+1}$ is an eigenvector of $\boldsymbol{X}^T\boldsymbol{X}$. Then show that the quantity in (1) is maximized when the $\boldsymbol{v}_{M+1}$ is the eigenvector with eigenvalue λ_{M+1}.

Problem 3 Naive Bayes (15 points)

Recall the naive Bayes model. Given a random variable $X \in R^D$ and a dependent class variable $Y \in [C]$, the joint distribution of features X and class Y is defined as

$$P(X = \mathbf{x}, Y = c) = P(Y = c)P(X = \mathbf{x}|Y = c) = P(Y = c)\prod_{d=1}^{D} P(X_d = x_d|Y = c)$$

In this problem, we consider a naive Bayes model where each feature x_d of each class c is modeled as a (separate) Gaussian. That is,

$$P(X_d = x_d \mid Y = c\,; \mu_{cd}, \sigma_{cd}) = \frac{1}{\sqrt{2\pi}\sigma_{cd}} \exp\left(-\frac{(x_d - \mu_{cd})^2}{2\sigma_{cd}^2}\right)$$

where μ_{cd} and σ_{cd} are the mean and the standard deviation, respectively. Moreover, we model Y as a multinomial distribution with parameter $\boldsymbol{\theta}$ (a distribution over C elements). That is,

$$P(Y = c\,; \boldsymbol{\theta}) = \theta_c \quad \forall c \in [C].$$

3.1 What are the parameters to be learned in this model? **(6 points)**

3.2 Given the dataset $\{(\mathbf{x}_n \in R^D, y_n \in [C])\}_{n=1}^{N}$, assumed to be drawn i.i.d. from this model, write down explicitly the expression for the joint log-likelihood. **(9 points)**