Tutorial 10 Solutions

Sunday 13th November, 2016

1 Gaussian Discriminant Analysis

1.1 Quadratic separating surface

- Consider the following Gaussian Discriminant Classifier discussed in class. Let us say we have K classes with a multivariate Gaussian Model $\mathcal{N}(\mu_i, \Sigma_i)$ fitted for each class:

$$P(\phi(x)|C_1) = \mathcal{N}(\mu_1, \Sigma_1)$$

$$P(\phi(x)|C_i) = \mathcal{N}(\mu_i, \Sigma_i)$$

$$P(\phi(x)|C_K) = \mathcal{N}(\mu_K, \Sigma_i)$$

- Assumption: $\phi(x)$ is generated using **exactly one** $\mathcal{N}(\mu_i, \Sigma_i)$
- In the case of $K = 2$, with $P(C_1)$ and $P(C_2)$ known, separating surface will be $\{\phi(x) \mid P(C_1|\phi(x)) = P(C_2|\phi(x))\}$.

Prove that the separating surface will be **quadratic**.

ANSWER:

- If $\phi(x) \sim \mathcal{N}(\mu_i, \Sigma_i)$ (where $\phi(x) \in \Re^m$) then

$$p(\phi(x) \mid C_i) = \frac{1}{(2\pi)^{\frac{m}{2}}|\Sigma_i|^{\frac{1}{2}}} exp\frac{-(\phi(x) - \mu_i)^T\Sigma_i^{-1}(\phi(x) - \mu_i)}{2}$$

- So, the separating surface is $\phi(x)$ such that $\{\phi(x) \mid P(C_1|\phi(x)) = P(C_2|\phi(x))\} \Longleftrightarrow \{\phi(x) \mid P(\phi(x) \mid C_1)P(C_1) = P(\phi(x) \mid C_2)P(C_2)\} \Longleftrightarrow$ after taking logs, $\phi(x)$ such that

$$-(\phi(x) - \mu_1)^T\Sigma_1^{-1}(\phi(x) - \mu_1) + (\phi(x) - \mu_2)^T\Sigma_2^{-1}(\phi(x) - \mu_2) = b$$

where b contains terms independent of $\phi(x)$.

- This is indeed a **QUADRATIC equation**!

1.2 Linear Discriminant Analysis

- Now consider the following variant of the above Gaussian Discriminant Classifier. Let us say we have K classes with a multivariate Gaussian Model $\mathcal{N}(\mu_i, \Sigma)$ fitted for each class. That is, the covariance matrix Σ is now shared across the classes:

$$P(\phi(x)|C_1) = \mathcal{N}(\mu_1, \Sigma)$$
$$P(\phi(x)|C_i) = \mathcal{N}(\mu_i, \Sigma)$$
$$P(\phi(x)|C_K) = \mathcal{N}(\mu_K, \Sigma)$$

- Assumption: $\phi(x)$ is generated using **exactly one** $\mathcal{N}(\mu_i, \Sigma)$.

- As before, in the case of $K = 2$, with $P(C_1)$ and $P(C_2)$ known, separating surface will be $\{\phi(x) \mid P(C_1|\phi(x)) = P(C_2|\phi(x))\}$.

1. Q: Prove that the separating surface will now be **linear**.

   Answer:

   - If $\phi(x) \sim \mathcal{N}(\mu_i, \Sigma)$ (where $\phi(x) \in \Re^m$) then

$$p(\phi(x) \mid C_i) = \frac{1}{(2\pi)^{\frac{m}{2}}|\Sigma|^{\frac{1}{2}}} exp\frac{-(\phi(x) - \mu_i)^T\Sigma^{-1}(\phi(x) - \mu_i)}{2}$$

   - So, the separating surface is $\phi(x)$ such that $\{\phi(x) \mid P(C_1|\phi(x)) = P(C_2|\phi(x))\}$ $\Longleftrightarrow$ $\{\phi(x) \mid P(\phi(x) \mid C_1)P(C_1) = P(\phi(x) \mid C_2)P(C_2)\}$ $\Longleftrightarrow$ after taking logs, $\phi(x)$ such that

$$-(\phi(x) - \mu_1)^T\Sigma^{-1}(\phi(x) - \mu_1) + (\phi(x) - \mu_2)^T\Sigma^{-1}(\phi(x) - \mu_2) = b$$

   where b contains terms independent of $\phi(x)$
   $\Longleftrightarrow$ ...the above expression can actually be simplified by canceling out he terms involving $\phi^T(x)\Sigma^{-1}\phi(x)$

$$-\phi^T(x)\Sigma^{-1}\phi(x) \ldots + \phi^T(x)\Sigma^{-1}\phi(x) = b$$

   to finally give

$$2\phi^T(x)\Sigma^{-1}\mu_1 + 2\phi^T(x)\Sigma^{-1}\mu_2 - \mu_1^T(x)\Sigma^{-1}\mu_1 - \mu_2^T(x)\Sigma^{-1}\mu_2 = b$$

   that is,

$$2\phi^T(x)\Sigma^{-1}\mu_1 + 2\phi^T(x)\Sigma^{-1}\mu_2 - \mu_1^T(x)\Sigma^{-1}\mu_1 = b'$$

   which is a **LINEAR equation**! Here, $b' = \mu_1^T(x)\Sigma^{-1}\mu_1 + \mu_2^T(x)\Sigma^{-1}\mu_2 + b$ is independent of $\phi(x)$.

2. Q: What will be the maximum likelihood estimates for μ_i and Σ in this new case of different means but shared covariance matrix?

   ANSWER: The Maximum Likelihood estimate for $\hat{\mu}_i$ will be the same as that for the Quadratic Discriminant Analysis, but that for a shared and single covariance estimate $\hat{\Sigma}$ will correspond to the average of covariance matrix estimates across examples from all the classes

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \phi(x_j^i)$$

Same as for QDA

Earlier for QDA

$$\hat{\Sigma}_i = \frac{1}{n_i} \sum_{x \in C_i} (\phi(x) - \mu_i)(\phi(x) - \mu_i)^T$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\phi(x_i^j) - \mu_i)(\phi(x_i^j) - \mu_i)^T$$

2 EM Algorithm for Mixture of Gaussians

Q: Show that the following algorithm for estimating the mean μ_i, the covariance matrix Σ_i and mixture components π_i for a mixture of Gaussians is an instance of the general EM algorithm

ANSWER: Has been discussed in Lecture 27.

Initialize $\mu_i^{(0)}$ to different random values and $\Sigma_i^{(0)}$ to I. Now iterate between the following **E Step** and **M Steps**:

E Step:

1. For the posterior $p(z_i \mid \phi(x_j), \mu, \Sigma)$

$$p^{(t+1)}(z_i \mid \phi(x_j), \theta) = \frac{\pi_i \mathcal{N}\left(\phi(x); \mu_i^{(t)}, \Sigma_i^{(t)}\right)}{\sum_{l=1}^{K} \pi_l \mathcal{N}\left(\phi(x); \mu_l^{(t)}, \Sigma_l^{(t)}\right)}$$

M Steps:

1. For the prior π_i

$$\pi_i^{(t+1)} = \frac{1}{n} \sum_{j=1}^{n} p^{(t+1)}(z_i \mid \phi(x_j), \theta)$$

2. For μ_i

$$\mu_i^{(t+1)} = \frac{\sum_{j=1}^{n} p^{(t+1)}(z_i \mid \phi(x_j), \theta) \phi(x_j)}{\sum_{j=1}^{n} p^{(t+1)}(z_i \mid \phi(x_j), \theta)}$$

3. For Σ_i

$$\Sigma_i^{(t+1)} = \frac{\sum_{j=1}^{n} p^{(t+1)}(z_i \mid \phi(x_j), \theta)\left(\phi(x_j) - \mu_i^{(t+1)}\right)\left(\phi(x_j) - \mu_i^{(t+1)}\right)^T}{\sum_{j=1}^{n} p^{(t+1)}(z_i \mid \phi(x_j), \theta)}$$

Q: Note that this algorithm is for the Mixture of Gaussians assuming a different covariance matrix Σ_i for each class C_i. What will be the algorithm like, if we assume a shared covariance matrix Σ across all classes (that is, the Linear Discriminant Analysis discussed in Section 1.2)?

ANSWER: We will simply build on the solution to the Linear Discriminant case from Section 2.1 and simply replace multiple class-specific estimates Σ_i with a single estimate Σ:

MStep: → $\Sigma^{(t+1)} = \frac{1}{m}\sum_x \sum_i q_i^t(z_i|x)(\phi(x) - \mu_i^{t+1})(\phi(x) - \mu_i^{t+1}) / \sum_x \sum_i q^t(z_i|x)$ (based on (*))

3 Convergence of Hard K-Means Algorithm

Prove the following claim: The K-Means Clustering algorithm will converge in a finite number of iterations.

1. **Proof Sketch:** At each iteration, the K-Means algorithm reduces the objective $\sum_{j=1}^{m}\sum_{l=1}^{K} P_{l,j}\|\phi\left(\mathbf{x}^{(j)}\right) - \mu_l\|^2$ and stops when this objective does not reduce any further.
2. Hint1: $P^{(t+1)} = \underset{P}{\operatorname{argmin}} \sum_{j=1}^{m}\sum_{l=1}^{K} P_{l,j}\|\phi\left(\mathbf{x}^{(j)}\right) - \mu_l^{(t)}\|^2$ (E step equivalent)
3. Hint2: $\mu^{(t+1)} = \underset{\mu}{\operatorname{argmin}} \sum_{j=1}^{m}\sum_{l=1}^{K} P_{l,j}^{(t+1)}\|\phi\left(\mathbf{x}^{(j)}\right) - \mu_l\|^2$ (M step equivalent)

① 2 must have reduced objective if any $P_{i,j}$ changed

4. Hint3: Only a finite number of combinations of $P_{l,j}$ are possible. (2) 2^{Km}

4 (Optional) Bayesian Inference from Multinomial to Naive Bayes (useful for text data)

First we summarize the conjugate prior, MLE and Bayesian estimation for Multinomial

- In the case of the Multinomial distribution (extension to the binomial distribution) a variable X could assume one of t possible values $V_1, V_2 \ldots V_t$ with parameters $\Pr(X = V_j) = \mu_j$. Each observation $\mathbf{X}_k$ (for $k \in [1, n]$) is modeled as a vector $\mathbf{X}_k = [X_{k,1} \ldots X_{k,j} \ldots X_{k,t}]$ with $X_{k,j} = 1$ if and only if value of $\mathbf{X}_k$ was observed to be V_j and $X_{k,j} = 0$ otherwise. Eg: In the case of the toss of dice, $t = 6$.

- The maximum likelihood estimate for the mean is given by:

$$\hat{\mu}_j = \frac{\sum_{k=1}^{n} X_{k,j}}{n} = \frac{n_j}{n} \tag{1}$$

  where, given n iid observations of a multinomial random variable X, $n_j = \sum_{k=1}^{n} X_{k,j}$ is the number of times $X = V_j$ was observed,

- Conjugate prior follows $\text{Dir}(\alpha_1 \dots \alpha_n)$

- Posterior is $\text{Dir}(\dots \alpha_l + \sum_{k=1}^{n} X_{k,j} \dots)$

- The expectation of μ under $\text{Dir}(\alpha_1 \dots \alpha_n)$ is given by:

$$E\left[\mu\right]_{Dir(\alpha_1 \dots \alpha_n)} = \left[\frac{\alpha_1}{\sum \alpha_l} \cdots \frac{\alpha_1}{\sum \alpha_l}\right] \tag{2}$$

- The (posterior) expectation of μ under $\text{Dir}(\dots \alpha_j + \sum_{k=1}^{n} X_{k,j} \dots)$ is given by:

$$E\left[\mu\right]_{Dir(\dots \alpha_j + \sum_{k=1}^{n} X_{k,j} \dots)} = \left[\frac{\alpha_1 + \sum_k X_{k,1}}{\sum \alpha_l + n} \cdots \frac{\alpha_j + \sum_k X_{k,j}}{\sum \alpha_l + n} \dots\right] \tag{3}$$

Now recall that we extended single Multinomial to Multinomial Naive Bayes that has class conditioned independent features, each of which is Multinomial. We also discussed Maximum Likelihood estimation of Multinomial Naive Bayes. All of that is summarized below:

- $< X_k, C_i >$: Tuple with example X_k belonging to class C_i. $\Pr(C_i)$ is prior probability of class C_i.

- $\phi_1(X_k), \dots, \phi_m(x_k)$: The feature vector for X_k

- $P(\phi_q(x)|C_i) \sim Mult(\mu^q_{1,i} \dots \mu^q_{t_q,i})$; that is, each feature ϕ_q follows multinomial distribution Bayes

  1. $[V^1_1 \dots V^1_{t_1}] \dots [V^q_1 \dots V^q_{t_q}] \dots [V^m_1 \dots V^m_{t_m}]$: Set of values that could be taken by each of $\phi_1, \phi_2 \dots \phi_m$ respectively
  2. $[\mu^1_{1,i} \dots \mu^1_{t_1,i}] \dots [\mu^q_{1,i} \dots \mu^q_{t_q,i}] \dots [\mu^m_{1,i} \dots \mu^m_{t_m,i}]$: Parameters for each of $\phi_1, \phi_2 \dots \phi_m$ respectively for class C_i

- $P(\phi_1(x) \dots \phi_m(x)|C_i) = \prod_{q=1}^{m} P(\phi_q(x)|C_i)$: Feature are independent given the class

- Maximum Likelihood Estimate for Naive Bayes: Let

$$\begin{aligned}
\#C_i = &\text{ No. of times } c(X_k) = C_i \text{ across all } k\text{'s in the dataset} \\
n^q_{j,i} = &\text{ No. of times } \phi_q(X_k) = V_j \text{ and } c(X_k) = C_i \text{ across all the } k\text{'s} \\
n^q_{j,i} = &\sum_k \delta\big(\phi_q(X_k), V^q_j\big)\delta\big(c(Xk_i), C_i\big)
\end{aligned}$$

then

$$\hat{\mu}^q_{j,i} = \frac{n^q_{j,i}}{\sum_{j'=1}^{n} n^q_{j',i}}$$
$$\widehat{Pr}_{c_i} = \frac{\#C_i}{\sum_{i'} \#C_{i'}}$$

Q: Extend Bayesian Inference using the Dirichlet prior from Multinomial to Naive Bayes

ANSWER: The setting for Naive Bayes with Dirichlet prior on Multivariate Bernoulli distribution is as follows

- For each data point X_k which belongs to class C_i there are a set of m features given by $\phi_1(X_k) \dots \phi_q(X_k) \dots \phi_m(X_k)|C_i$
- Each parameter (vector) $\mu^q_i = [\mu^q_{1,i} \dots \mu^q_{j,i} \dots \mu^q_{t_q,i}]$ corresponding to a feature $\phi_q(.)$ has a probability distribution given by
  $p(\mu^1_i) \sim Dir(a^1_{1,i} \dots a^1_{j,i} \dots a^1_{t_1,i}) \dots$
  $p(\mu^q_i) \sim Dir(a^q_{1,i} \dots a^q_{j,i} \dots a^q_{t_q,i}) \dots$
  $p(\mu^m_i) \sim Dir(a^m_{1,i} \dots a^m_{j,i} \dots a^m_{t_m,i})$
  where $a^q_{j,i}$ are the multivariate dirichlet prior parameters
- Let $n^q_{j,i}$ be the number of times attribute $\phi_q(.)$ was observed with value V^q_j amongst examples belonging to class C_i.
- Then, from our previous analysis of Bayesian estimation for Multinomial[1],
  $p(\mu^1_i \mid D) \sim Dir(a^1_{1,i} + n^1_{1,i} \dots a^1_{j,i} + n^1_{j,i} \dots a^1_{t_1,i} + n^1_{t_1,i}) \dots$
  $p(\mu^q_i \mid D) \sim Dir(a^q_{1,i} + n^q_{1,i} \dots a^q_{j,i} + n^q_{j,i} \dots a^q_{t_q,i} + n^q_{t_q,i}) \dots$
  $p(\mu^m_i \mid D) \sim Dir(a^m_{1,i} + n^m_{1,i} \dots a^m_{j,i} + n^m_{j,i} \dots a^m_{t_m,i} + n^m_{t_m,i})$.

[1] See page 12 of http://23.253.82.180/course/307/865/2238.