

Tutorial 10

Thursday 3rd 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**.

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**.
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?

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

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 $p(z_i | \phi(x_j), \mu, \Sigma)$

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

M Steps:

1. For the prior π_i

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

2. For μ_i

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

3. For Σ_i

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

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 2.1)?

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 $\sum_{j=1}^m \sum_{l=1}^K P_{l,j} \|\phi(\mathbf{x}^{(j)}) - \mu_l\|^2$ and stops when this objective does not reduce any further.
2. Hint1: $P^{(t+1)} = \operatorname{argmin}_P \sum_{j=1}^m \sum_{l=1}^K P_{l,j} \|\phi(\mathbf{x}^{(j)}) - \mu_l^{(t)}\|^2$
3. Hint2: $\mu^{(t+1)} = \operatorname{argmin}_\mu \sum_{j=1}^m \sum_{l=1}^K P_{l,j}^{(t+1)} \|\phi(\mathbf{x}^{(j)}) - \mu_l\|^2$
4. Hint3: Only a finite number of combinations of $P_{i,j}$ are possible.

4 Bayesian Inference from Multinomial to Naive Bayes

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 $V_1, V_2 \dots V_t$ with parameters $\Pr(X = V_j) = \mu_j$. Each observation \mathbf{X}_k (for $k \in [1, n]$) is modeled as a vector $\mathbf{X}_k = [X_{k,1} \dots X_{k,j} \dots X_{k,t}]$ 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:

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

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 $\operatorname{Dir}(\alpha_1 \dots \alpha_n)$
- Posterior is $\operatorname{Dir}(\dots \alpha_l + \sum_{k=1}^n X_{k,j} \dots)$
- The expectation of μ under $\operatorname{Dir}(\alpha_1 \dots \alpha_n)$ is given by:

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

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

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

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:

- $\langle X_k, C_i \rangle$: 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
- $P(\phi_q(x)|C_i) \sim Mult(\mu_{1,i}^q \dots \mu_{t_q,i}^q)$; that is, each feature ϕ_q follows multinomial distribution Bayes
 1. $[V_1^1 \dots V_{t_1}^1] \dots [V_1^q \dots V_{t_q}^q] \dots [V_1^m \dots V_{t_m}^m]$: Set of values that could be taken by each of $\phi_1, \phi_2 \dots \phi_m$ respectively
 2. $[\mu_{1,i}^1 \dots \mu_{t_1,i}^1] \dots [\mu_{1,i}^q \dots \mu_{t_q,i}^q] \dots [\mu_{1,i}^m \dots \mu_{t_m,i}^m]$: 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

$\#C_i$ = No. of times $c(X_k) = C_i$ across all k 's in the dataset

$n_{j,i}^q$ = No. of times $\phi_q(X_k) = V_j$ and $c(X_k) = C_i$ across all the k 's

$$n_{j,i}^q = \sum_k \delta(\phi_q(X_k), V_j^q) \delta(c(X_k), C_i)$$

then

$$\hat{\mu}_{j,i}^q = \frac{n_{j,i}^q}{\sum_{j'=1}^n n_{j',i}^q}$$

$$\widehat{Pr}_{c_i} = \frac{\#C_i}{\sum_{i'} \#C_{i'}}$$

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