Advanced Machine Learning - Homework 1

宁雪妃

March 6, 2017

1 Problem 1

a **Consider a real, symmetric matrix Σ whose eigenvalue equation is given by:**

$$\Sigma u_i = \lambda_i u_i$$

By taking the complex conjugate of this equation and subtracting the original equation, and then forming the inner product with eigenvector u_i, show that the eigenvalues λ_i are real. Similarly, use the symmetry property of Σ to show that two eigenvectors u_i and u_j will be orthogonal provided $\lambda_j \neq \lambda_i$. Finally, show that without loss of generality, the set of eigenvectors can be chosen to be orthonormal, even if some of the eigenvalues are zero.

To prove all eigenvalues λ_i of a symmetric real matrix are real, we left multiply the conjugate-transpose vector of u_i: $u_i^{'}$.

$$u_i^{'}\Sigma u_i = \lambda_i u_i^{'} u_i$$

Because Σ is a real symmetric matrix, its conjugate transpose matrix is itself: $\Sigma = \Sigma^{'}$, so, we have$u_i^{'}\Sigma u_i = (\Sigma u_i)' u_i = \lambda_i^* u_i' u_i$. Compared with the equation above, as $u_i \neq 0 \rightsquigarrow u_i' u_i > 0$, we have $\lambda_i = \lambda_i^*$, which means all λ_i is real.

To prove the eigen vectors corresponding to different eigenvalues of a real symmetric matrix is orthogonal, assume u_i, u_j is the corresponding eigen vector of two different eigen-

values $\lambda_i \neq \lambda_j$ of Σ, we have:

$$\begin{aligned} \Sigma u_i &= \lambda_i u_i \\ \Sigma u_j &= \lambda_j u_j \end{aligned} \tag{1.1}$$

Left multiply u_j^T to the first eigen equation:

$$\begin{aligned} u_j^T \lambda_i u_i = u_j^T \Sigma u_i = (\Sigma u_j)^T u_i = \lambda_j u_j^T u_i \\ \rightsquigarrow (\lambda_i - \lambda_j) u_j^T u_i = 0 \end{aligned} \tag{1.2}$$

As $\lambda_i \neq \lambda_j$, we have $u_j^T u_i = 0$, proved.

According to schur theorem: any matrix whose eigenvalues are all real is orthogonal similar to a upper triangular matrix. In the meantime, Σ is symmetric, so that upper triangular matrix is a diagonal matrix. So, Σ is orthogonal similar to a diagonal matrix, which means the every column of the orthogonal matrix is a eigenvector of Σ, and they are orthogonal.

This proposition can be illustrated in another way, as long as the eigenvectors corresponding to the same eigenvalue can be orthogonalized, while we have already proved eigenvectors corresponding to different eigenvalues are orthogonal, we can say the eigenvectors set can be choosen so that every two eigenvectors are orthogonal. There is a theorem that tells us: for a real symmetric matrix, the eigen space of a r-algebraic-multiplicity eigenvalue must be a r-dim space. So we can take any orthogonal basis in thir r-dim space as the eigen vectors corresponding to this eigenvalue.

b **Refer to slides about PCA, where we perform eigen-decomposition on**

$$A = \frac{1}{N} \sum_{1=1}^{N} x_i x_i^T$$

Prove A is a symmetric and positive semi-definite matrix.

Prove A is symmetric:

$$A^T = \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T = A$$

Prove A is positive semi-definite:

$$\forall y, y^T A y = \frac{1}{N} \sum_{i=1}^{N} (x_i^T y)^T (x_i^T y) \geq 0$$

2 Problem 2

Given a set of i.i.d data $X = \{x_1, \ldots, x_N\}$ drawn from $N(x; \mu, \Sigma)$, we want to estimate (μ, Σ) by MLE.

a **Write the log likelyhood function.**

$$\begin{aligned} L(X) = P(\{x_1,\ldots,x_N\}|\mu,\Sigma) &= \prod_{i=1}^{N} P(x_i|\mu,\Sigma) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \exp(-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)) \\ &= \frac{1}{(2\pi)^{\frac{nN}{2}}\Sigma^{\frac{N}{2}}} \exp(-\frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)) \\ \text{logL}(X) = \log(L(X)) &= -\frac{N}{2}\log(|\Sigma|) - \frac{Nn}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) \end{aligned}$$

b **Take the derivative of log likelyhood function w.r.t μ, show that**

$$\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\frac{\partial \text{logL(X)}}{\partial \mu} = \sum_{i=1}^{N}\Sigma^{-1}(x_i-\mu) = \Sigma^{-1}\sum_{i=1}^{N}(x_i-\mu) = 0$$

As Σ^{-1} is non-singular, its null space is empty set. So, we must have:

$$\sum_{i=1}^{N}(x_i-\mu) = 0 \rightsquigarrow \mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

c **Take the derivative of log likelihood function w.r.t. Σ, show that**

$$\sum_{ML} = \frac{1}{N}\sum_{i=1} N(x_i-\mu_{ML})(x_i-\mu_{ML})^T$$

$$\frac{\partial \text{logL(X)}}{\partial \Sigma} = -\frac{N}{2}\Sigma^{-T} + \frac{1}{2}\sum_{i=1}^{N}\Sigma^{-T}(x_i-\mu)(x_i-\mu)^T\Sigma^{-T} = \frac{\Sigma^{-T}}{2}(\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T\Sigma^{-T} - N) = 0$$

As Σ^{-1} is non-singular, we have $\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T\Sigma^{-T} = 1$. Right multiply Σ to the equation, and use the property $\Sigma^T = \Sigma$, we get:

$$\Sigma_{ML} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{ML})(x_i-\mu_{ML})^T$$

d **Evaluate expectations of μ_{ML} and Σ_{ML}, show μ_{ML} is unbiased but Σ_{ML} is biased.**

$$E(\mu_{ML}) = \frac{1}{N}\sum_{i=1}^{N}E(x_i) = \frac{1}{N}\sum_{i=1}^{N}\mu = \mu$$

μ_{ML} is an unbiased estimation.

$$\begin{aligned} E(\Sigma_{ML}) &= \frac{1}{N}\sum_{i=1}^{N}E(x_i x_i^T + {\mu_{ML}}^2 - \mu_{ML}x_i^T - x_i\mu_{ML}^T) = \Sigma + E(\mu_{ML}\mu_{ML}^T) - \frac{1}{N}E(\sum_{i=1}^{N}\mu_{ML}x_i^T) - \frac{1}{N}E(\sum_{i=1}^{N}x_i\mu_{ML}^T) \\ &= \Sigma - E(\mu_{ML}\mu_{ML}^T) \neq \Sigma \end{aligned}$$

Σ_{ML} is a biased estimation.

3 Problem 3

For support vector machines, the class-conditional distributions may overlap, we therefore modify the support vector machine so as to allow some of the training points to be misclassified. For un-separable case, the formalization of the optimal problem becomes: Given $\{x_i, y_i\}, i = 1,\ldots,N, y_i \in \{1,1\}$ **are training examples,**

$$\begin{aligned} \min_{\omega,b} \frac{\|w\|^2}{2} + C\sum_{i=1}^{N}\xi_i \\ \text{s.t. } y^{(i)}(\omega^T x^{(i)} + b) \geq 1 - \xi_i, 1 \leq i \leq N \\ \xi_i \geq 0, 1 \leq i \leq N \end{aligned}$$

where the ξ_i **denotes the slack variable penalty, and the parameter** C **controls the trade-off betwwen the slack variable penalty and the margin. Please give the solutions of** ω **and** b**.**

a **Give the corresponding Lagrangian and the set of KKT conditions.** Lagrangian:

$$l(\omega, b, \{\xi_i\}, \{a_i\}, \{\beta_i\}) = \frac{\|\omega\|^2}{2} + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N} a_i(y^{(i)}(w^T x^{(i)} + b) + \xi_i - 1) - \sum_{i=1}^{N}\beta_i\xi_i$$

Corresponding KKT condition of this primal problem:

$$\begin{aligned} a_i(y^{(i)}(w^T x^{(i)} + b) + \xi_i - 1) = 0 \\ y^{(i)}(w^T x^{(i)} + b) + \xi_i - 1 \geq 0 \\ \xi_i \geq 0 \\ a_i \geq 0 \\ \beta_i \geq 0 \\ \text{for } i = 1,\ldots,N \end{aligned}$$

b **Optimize out** ω**,** b **and** $\{\xi_i\}$**.** The original objective function is equal to

$$\max_{\{a_i\},\{\beta_i\}} (l(\omega, b, \{\xi_i\}, \{a_i\}, \{\beta_i\}))$$

in the feasible solution space. As the primal problem satisfies the regularity conditions of strong duality, so the optimal solution of the dual problem

$$\max_{\{a_i\},\{\beta_i\}} \min_{\omega,b,\{\xi_i\}} l(\omega, b, \{\xi_i\}, \{a_i\}, \{\beta_i\})$$

is the same as the optimal solution of the primal problem (the duality gap is 0). So, we just need to solve the dual problem, after which we can solve ω and b according to the optimal

solution of the inner minimizing optimization problem: $\omega, b = f^*(\{a_i\}, \{\beta_i\})$. The optimal solution of the inner minimizing optimization problem must satisfies:

$$\frac{\partial l(\omega, b, \{\xi_i\}, \{a_i\}, \{\beta_i\})}{\partial \omega} = \omega - \sum_{i=1}^{N} a_i y^{(i)} x^{(i)} = 0$$
$$\frac{\partial l(\omega, b, \{\xi_i\}, \{a_i\}, \{\beta_i\})}{\partial b} = -\sum_{i=1}^{N} a_i y^{(i)} = 0$$
$$\frac{\partial l(\omega, b, \{\xi_i\}, \{a_i\}, \{\beta_i\})}{\partial \xi_i} = C - a_i - \beta_i = 0$$

Substitute $\omega = \sum_{i=1}^{N} a_i y^{(i)} x^{(i)}$, $\sum_{i=1}^{N} a_i y^{(i)} = 0$ and $C - a_i - \beta_i = 0$ into the lagrangian, we get:

$$\begin{aligned} l(\{a_i\}, \{\beta_i\}) &= \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j y^{(i)} y^{(j)} x^{(i)} x^{(j)} + C \sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} a_i - \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j y^{(i)} y^{(j)} x^{(i)} x^{(j)} - \sum_{i=1}^{N} a_i \xi_i - \sum_{i=1}^{N} \beta_i \xi_i \\ &= -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j y^{(i)} y^{(j)} x^{(i)} x^{(j)} + \sum_{i=1}^{N} a_i \end{aligned}$$

Notice there do not exist β_i in this objective, however, the constraint of $\beta_i \geq 0$ together with the inner minmizing optimal condition $C - a_i - \beta_i = 0$ bring an additional constraint for a_i: $a_i \leq C$.

c **Give the dual Lagrangian.** Dual Lagrangian is:

$$l(\{a_i\}, \{\beta_i\}) = \min_{\omega, b, \{\xi_i\}} l(\omega, b, \{\xi_i\}, \{a_i\}, \{\beta_i\}) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j y^{(i)} y^{(j)} x^{(i)} x^{(j)} + \sum_{i=1}^{N} a_i$$

The dual problem is to maximize the dual lagrangian satisfying the following constraints:

$$0 \leq a_i \leq C, 1 \leq i \leq N$$
$$\sum_{i=1}^{N} a_i y_i = 0$$

d **Give the final solution for ω and the numerically stable solution of b.** After we found the optimal $\{a_i\}$ using some QP solver, we can easily construct $\omega^* = \sum_{i=1}^{N} a_i y^{(i)} x^{(i)}$. To calculate b, we just need to find a data point $(x^{(j)}, y^{(j)})$ that has $a_j > 0$ and $a_j \neq C$(this data point is a supporting but not erroneous point), we have $y^{(j)}(\omega^* x^{(j)} + b) - 1 = 0 \rightsquigarrow b = y^{(j)} - \omega^* x^{(j)}$.