CS 475 Machine Learning: Homework 2
Supervised Classifiers 1,
Probability, Linear Algebra and Decision Trees

Due: Tuesday September 29, 2015, 11:59pm

100 Points Total Version 1.0

Li-Yi Lin / llin34@jhu.edu

1 Analytical (50 points)

1) Fisher Linear Discriminant and Logistic Regression Classifiers (15 points) Generative models and discriminative models are somehow connected given certain scenarios. Suppose that we have samples from two classes with equal prior. The first class of samples have their features independent generated from a multivariate normal distribution $N(\mu_1, \Sigma)$, and the second class of samples have their features independently generated from a multivariate normal distribution $N(\mu_2, \Sigma)$.

(a) Prove that the class label y conditioning on the feature vector X follows a logistic regression model.

Ans:

We first use Bayes rule on $P(y|X)$:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

Since the two classes has equal prior, we have $P(y) = 1/2$ for each y. In addition, $P(X)$ can be represented as $\sum_y P(X|y)P(y)$. So, the original equation can be changed as below:

$$P(y|X) = \frac{P(X|y)P(y)}{\sum_y P(X|y)P(y)}$$

$$= \frac{\frac{1}{\sqrt{(2\pi)^k \Sigma}} e^{-\frac{1}{2}(X-\mu_1)^T \Sigma^{-1}(X-\mu_1)} \times \frac{1}{2}}{\frac{1}{\sqrt{(2\pi)^k \Sigma}} e^{-\frac{1}{2}(X-\mu_1)^T \Sigma^{-1}(X-\mu_1)} \times \frac{1}{2} + \frac{1}{\sqrt{(2\pi)^k \Sigma}} e^{-\frac{1}{2}(X-\mu_2)^T \Sigma^{-1}(X-\mu_2)} \times \frac{1}{2}}$$

$$= \frac{e^{-\frac{1}{2}(X-\mu_1)^T \Sigma^{-1}(X-\mu_1)}}{e^{-\frac{1}{2}(X-\mu_1)^T \Sigma^{-1}(X-\mu_1)} + e^{-\frac{1}{2}(X-\mu_2)^T \Sigma^{-1}(X-\mu_2)}}$$

$$= \frac{1}{1 + e^{-\frac{1}{2}(X-\mu_2)^T \Sigma^{-1}(X-\mu_2) + \frac{1}{2}(X-\mu_1)^T \Sigma^{-1}(X-\mu_1)}}$$

$$= \frac{1}{1 + e^{-\frac{1}{2}(X^T \Sigma^{-1} X - X^T \Sigma^{-1} \mu_2 - \mu_2^T \Sigma^{-1} X + \mu_2^T \Sigma^{-1} \mu_2 - X^T \Sigma^{-1} X + X^T \Sigma^{-1} \mu_1 + \mu_1^T \Sigma^{-1} X - \mu_1^T \Sigma^{-1} \mu_1)}}$$

Since $X^T\Sigma^{-1}\mu_1$ and $X^T\Sigma^{-1}\mu_2$ are numbers, we can transpose it. (Σ is a symmetric matrix)
$(X^T\Sigma^{-1}\mu_1)^T = \mu_1^T(\Sigma^{-1})^T X = \mu_1^T(\Sigma^T)^{-1}X = \mu_1^T\Sigma^{-1}X$
$(X^T\Sigma^{-1}\mu_2)^T = \mu_2^T(\Sigma^{-1})^T X = \mu_2^T(\Sigma^T)^{-1}X = \mu_2^T\Sigma^{-1}X$
So the equation becomes:

$$= \frac{1}{1+\mathrm{e}^{-\frac{1}{2}(-2\mu_2^T\Sigma^{-1}X+2\mu_1^T\Sigma^{-1}X+\mu_2^T\Sigma^{-1}\mu_2-\mu_1^T\Sigma^{-1}\mu_1)}}$$

$$= \frac{1}{1+\mathrm{e}^{-(\mu_1^T-\mu_2^T)\Sigma^{-1}X-\frac{1}{2}(\mu_2^T\Sigma^{-1}\mu_2-\mu_1^T\Sigma^{-1}\mu_1)}}$$

Although the equation has one additional fixed scalar, $-\frac{1}{2}(\mu_2^T\Sigma^{-1}\mu_2 - \mu_1^T\Sigma^{-1}\mu_1)$, at the power of e, it still follows a logistic regression model. Thus, we have proved that the class label y conditioning on the feature vector X follows a logistic regression model.

(b) Prove that the classifier based on the logistic regression model obtained in (a) is equivalent the optimal Fisher linear discriminant classifier. The optimal Fisher linear discriminant classifier is obtained using the population means and covariance matrix; see section 4.1.4 in Bishop.

Hint: You only need to show that both classifiers use the same decision rule.
Ans:
The logistic regression model obtained in (a) uses the population means and covariance matrix as its parameter in the sigmoid function, and classifies an instance by its sigmoid value according to a decision boundary. For the optimal Fisher linear discriminant classifier, it also use its population means and covariance matrix to find its parameters. The function is shown below:

$$\mathbf{w} \propto S_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$

where $\mathbf{w}$ is its parameter vector, S_W^{-1} is the within-class covariance matrix, and $\mathbf{m}_1$ and $\mathbf{m}_2$ are the mean vectors of two classes. When performing classification task, the optimal Fisher linear discriminant classifier uses the following function to find projected value:

$$y = f(\mathbf{w}^T X)$$

where X is a feature vector of an instance. And it classifies the instance by comparing y with a decision boundary. In the logistic regression model in (a), the decision boundary is 0.5 for whole function

$$\frac{1}{1+\mathrm{e}^{-(\mu_1^T-\mu_2^T)\Sigma^{-1}X-\frac{1}{2}(\mu_2^T\Sigma^{-1}\mu_2-\mu_1^T\Sigma^{-1}\mu_1)}}$$

And that means when $(-(\mu_1^T - \mu_2^T)\Sigma^{-1}X - \frac{1}{2}(\mu_2^T\Sigma^{-1}\mu_2 - \mu_1^T\Sigma^{-1}\mu_1))$ is larger than zero, the logistic sigmoid value will be larger than 0.5. When it is small than zero, its logistic sigmoid value will be less than 0.5. In other words, we can use zero as a decision boundary to the function $(-(\mu_1^T - \mu_2^T)\Sigma^{-1}X - \frac{1}{2}(\mu_2^T\Sigma^{-1}\mu_2 - \mu_1^T\Sigma^{-1}\mu_1))$. For the optimal Fisher linear discriminant classifier also use zero as a decision boundary to check the value of y. Thus, these two classifiers uses the same decision rule.

2) Linear Models (10 points) Besides the least square estimators, machine learning researchers are also interested in another type of estimators – maximum likelihood estimators. Consider a linear model $y = X\beta+\epsilon$, where $X \in \mathbb{R}^{n\times d}$ is the design matrix, $y \in \mathbb{R}^n$ is the response vector, and $\epsilon \in \mathbb{R}^n$ is the random noise with each entry independently sampled from $N(0, \sigma^2)$. Please derive the maximum likelihood estimator of β and σ.

Ans:

Since each entry of ϵ is sampled from $N(0, \sigma^2)$, every $y_i \in y$ follows $N(\sum_{j=1}^{n} x_{ij}\beta_j, \sigma^2)$ and is independent from each other. We can get the probability of y by following equation:

$$P(y) = \prod_{i=1}^{n} N(y_i|\sum_{j=1}^{d} x_{ij}\beta_j, \sigma^2)$$

$$P(y) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(y_i-\sum_{j=1}^{d} x_{ij}\beta_j)^2}$$

To solve the above equation, we apply nature log on both sides.

$$\ln P(y) = \sum_{i=1}^{n} \{-\frac{1}{2}\ln 2\pi - \ln\sigma - \frac{1}{2\sigma^2}(y_i - \sum_{j=1}^{d} x_{ij}\beta_j)^2\}$$

We want to maximize the log-likelihood regarding β and σ. First we have to find every β_j such that $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{d} x_{ij}\beta_j)^2$ is minimized. We solve it by following equation:

$$\frac{\partial}{\partial\beta_j}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{d} x_{ij}\beta_j)^2 = 2\sum_{i=1}^{n}\{(y_i - \sum_{j=1}^{d} x_{ij}\beta_j)x_{ij}\} = 0$$

Since $\sum_{i=1}^{n}\{(y_i - \sum_{j=1}^{d} x_{ij}\beta_j)x_{ij}\}$ is actually a matrix multiplication. So we change it back to matrix form and solve it:

$$(y - X\beta)^T X = 0$$
$$(y^T - (X\beta)^T)X = 0$$
$$y^T X - \beta^T X^T X = 0$$
$$\beta^T X^T X = y^T X$$
$$\beta^T = y^T X(X^T X)^{-1}$$
$$\beta = (y^T X(X^T X)^{-1})^T$$
$$\beta = (X^T X)^{-1} X^T y$$

So we have find $\beta = (X^T X)^{-1} X^T y$.

To find the σ that makes the $\ln P(y)$ biggest, we apply partial derivative on σ and set it equal to 0. We solve it by following step:

$$\frac{\partial \ln P(y)}{\partial\sigma} = \sum_{i=1}^{n}\{-\frac{1}{\sigma} + \frac{2}{2\sigma^3}(y_i - \sum_{j=1}^{d} x_{ij}\beta_j)^2\} = 0$$

$$-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i-\sum_{j=1}^{d}x_{ij}\beta_j)^2=0$$

$$-\frac{n}{\sigma}+\frac{1}{\sigma^3}(y-X\beta)^T(y-X\beta)=0$$

$$n\sigma^2=(y-X(X^TX)^{-1}X^Ty)^T(y-X(X^TX)^{-1}X^Ty)$$

$$n\sigma^2=(y^T-y^TX(X^TX)^{-1}X^T)(y-X(X^TX)^{-1}X^Ty)$$

$$n\sigma^2=y^Ty-y^TX(X^TX)^{-1}X^Ty-y^TX(X^TX)^{-1}X^Ty+y^TX(X^TX)^{-1}X^TX(X^TX)^{-1}X^Ty$$

$$n\sigma^2=y^Ty-y^TX(X^TX)^{-1}X^Ty$$

$$\sigma=\sqrt{\frac{y^Ty-y^TX(X^TX)^{-1}X^Ty}{n}}$$

σ must be larger or equal to 0 since it is standard deviation.

3) Regularization and Overfitting. (5 points) Statisticians love linear models because these models are very simple and interpretable. Many variants of linear models has been proposed, and most of them are formulated as (penalized) least squares program. Here we have three least squares programs,

$$\hat{\beta}_0=\underset{\beta_0}{\operatorname{argmin}}\,\|y-X_1\beta_0\|_2^2, \tag{1}$$

$$(\hat{\beta}_1,\hat{\beta}_2)=\underset{\beta_1,\beta_2}{\operatorname{argmin}}\,\|y-X_1\beta_1-X_2\beta_2\|_2^2, \tag{2}$$

$$\hat{\beta}_3=\underset{\beta_3}{\operatorname{argmin}}\,\|y-X_1\beta_3\|_2^2+\lambda\|\beta_3\|_2^2, \tag{3}$$

where $\lambda>0$, $y\in\mathbb{R}^n$, $X_1\in\mathbb{R}^{n\times d_1}$, and $X_2\in\mathbb{R}^{n\times d_1}$. (3) is well known as the ridge regression. The square norm acts as a penalty function to reduce overfitting. Prove

$$\|y-X_1\hat{\beta}_3\|_2^2\geq\|y-X_1\hat{\beta}_0\|_2^2\geq\|y-X_1\hat{\beta}_1-X_2\hat{\beta}_2\|_2^2. \tag{4}$$

Ans:
We first prove that $\|y-X_1\hat{\beta}_3\|_2^2\geq\|y-X_1\hat{\beta}_0\|_2^2$. Since $\hat{\beta}_0$ is the answer vector that minimizes the form $\|y-X_1\beta\|_2^2$, any other vector β' in this form will not make it smaller than $\|y-X_1\hat{\beta}_0\|_2^2$ and so does $\hat{\beta}_3$. Therefore, we have proved that $\|y-X_1\hat{\beta}_3\|_2^2\geq\|y-X_1\hat{\beta}_0\|_2^2$.

Next, we need to prove that $\|y-X_1\hat{\beta}_0\|_2^2\geq\|y-X_1\hat{\beta}_1-X_2\hat{\beta}_2\|_2^2$. Since we already know that $\hat{\beta}_1$ and $\hat{\beta}_2$ is the answer vectors that minimize $\|y-X_1\beta_1-X_2\beta_2\|_2^2$, any other (β_1,β_2) pair for this form will not make it smaller. We set $\beta_1=\hat{\beta}_0$ and $\beta_2=\vec{0}$, then we get $\|y-X_1\hat{\beta}_0\|_2^2=\|y-X_1\hat{\beta}_0-X_2\vec{0}\|_2^2\geq\|y-X_1\hat{\beta}_1-X_2\hat{\beta}_2\|_2^2$. Thus, we have proved that $\|y-X_1\hat{\beta}_0\|_2^2\geq\|y-X_1\hat{\beta}_1-X_2\hat{\beta}_2\|_2^2$.

Finally, we combine the above two proofs, we have proved $\|y-X_1\hat{\beta}_3\|_2^2\geq\|y-X_1\hat{\beta}_0\|_2^2\geq\|y-X_1\hat{\beta}_1-X_2\hat{\beta}_2\|_2^2$.

4) Decision Tree (10 points) Let's investigate how accurately decisions trees can learn. We start by constructing a unit square ($[0; 1] \times [0; 1]$). We select n samples from the square, each with a binary label ($+1$ or -1), such that no two samples share either x or y coordinates. Unlike the programming above, each feature can be used multiple times in a decision tree. At each node we can only conduct a binary threshold split using one single feature.

(a) Prove that we can find a decision tree of depth at most $\log_2 n$, which perfectly labels all n samples.
Ans:
Since all the samples share no x or y coordinates, if we draw a vertical or horizontal line through a sample, there will not be another sample on the same line. In other words, we can always find a line that can equally divide the samples into two groups (assume n is a even number). Each feature can be used multiple times. For every procedure, we divide the samples into two group with the same size (assume the sample size is even). We repeat this procedure until we divide the samples into n labels. When we add one depth to the tree, dividing the samples into two groups, the group size will be half of the previous group size. Thus, the number of depth in a decision tree will be:

$$2^{depth} = n$$

$$depth = \log_2 n$$

Therefore, we have proved that we can find a decision tree of depth at most $\log_2 n$ that perfectly labels all n samples.

(b) If the samples can share either x or y coordinates but not both, can we still learn a decision tree which perfectly labels all n samples? Why or why not?
Ans:
Assume we have four samples. Label $+1$ at (0, 0), (1, 0), (0.5, 1) and Label -1 at (0.5, 0). In this situation, we can not construct a decision tree with depth of $\log_2 4 = 2$ such that the four samples can be perfectly labelled.

-1 what is depth of tree that perfectly labels?

5) Conjugate Prior (10 points) The conjugate priors are very popular in Bayesian data analysis. The formal description of the conjugate priors can be found in Chapter 2.4.2. of Bishop's PRML.

(a) Prove that the Gamma distribution with parameters α and β is a conjugate prior of the Poisson distribution with parameter λ.
Ans:
Let x_i follow Poisson distribution with parameter λ, then we have

$$x_i \sim Poisson(\lambda) => p(x_i|\lambda) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \propto \lambda^{x_i} e^{-\lambda}$$

and assume λ follows Gamma distribution with parameters α and β

$$\lambda \sim Gamma(\alpha, \beta) => P(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \propto \lambda^{\alpha-1} e^{-\beta\lambda}$$

Probability $P(X|\lambda)$ can be calculated as

$$P(X|\lambda) = \frac{\prod_{i=1}^{n} \lambda^{x_i} \mathrm{e}^{-\lambda}}{x_i!} = \frac{\lambda^{x_1+x_2+\ldots+x_n} \mathrm{e}^{-n\lambda}}{\prod_{i=1}^{n} x_i!} \propto \lambda^{x_1+x_2+\ldots+x_n} \mathrm{e}^{-n\lambda} = \lambda^{n\bar{X}} \mathrm{e}^{-n\lambda}$$

where $\sum_{i=1}^{n} x_i = n\bar{x}$ and X is a vector of x_i

By Bayes' theorem, we know

$$P(\lambda|X) = \frac{P(X|\lambda) \times P(\lambda)}{P(X)} \propto P(X|\lambda) \times P(\lambda)$$

Then we can get:

$$P(X|\lambda) \times P(\lambda) \propto \lambda^{n\bar{x}} \mathrm{e}^{-n\lambda} \times \lambda^{\alpha-1} \mathrm{e}^{-\beta\lambda} = \lambda^{n\bar{x}+\alpha-1} \mathrm{e}^{-(n+\beta)\lambda}$$

So we have

$$P(\lambda|x) \propto \lambda^{n\bar{x}+\alpha-1} \mathrm{e}^{-(n+\beta)\lambda} \sim Gamma(n\bar{x} + \alpha - 1, n + \beta)$$

Therefore, we have proved that Gamma distribution with parameters α and β is a conjugate prior of the Poisson distribution with parameter λ.

(b) Prove that the Beta distribution with parameters α and β is a conjugate prior of the geometric distribution with parameter p.
Ans:
Let x_i follow geometric distribution, then we have

$$x_i \sim Geometric(p) => P(x|P) = (1-p)^{x-1} p$$

And assume its parameter, p, follows Beta distribution, then we have

$$p \sim \text{Beta}(\alpha, \beta) => P(p|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1} (1-p)^{\beta-1} \propto p^{\alpha-1} (1-p)^{\beta-1}$$

By Bayes' theorem, we know posterior $\propto$ prior $\times$ likelihood. So we have

$$\text{posterior} \propto \prod_{i=1}^{n} \{(1-p)^{x_i-1} p\} \times p^{\alpha-1} (1-p)^{\beta-1}$$

Let $\frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$. Then above equation becomes

$$(1-P)^{n\bar{x}-n} p^n \times p^{\alpha-1} (1-p)^{\beta-1} = p^{n+\alpha-1} (1-p)^{n\bar{x}+\beta-n-1}$$

So, the posterior $\sim$ Beta$(n + \alpha, n\bar{x} + \beta - n)$. Thus we have proved that the Beta distribution with parameters α and β is a conjugate prior of geometric distribution with parameter p.