Linear Classification

Weikai Xu

School of Information Science and Technology
ShanghaiTech University

Mar 15th, 2020

Outline

Outline

Linear Regression of an indicator matrix

1. Compute the fitted output

$$\hat{f}(x) = \hat{B}\begin{pmatrix}1\\ x\end{pmatrix} = \begin{pmatrix}\hat{f}_1(x)\\ \hat{f}_2(x)\\ \vdots\\ \hat{f}_k(x)\end{pmatrix}$$

2. Classify x according to

$$\hat{G}(x) = \arg\max \hat{f}_k(x)$$

3. The above formula is equal to

$$\hat{G}(x) = \arg\min ||\hat{f}_k(x) - t_k||_2^2$$

where $t_k = (0, \cdots, 0, 1, 0, \cdots, 0)^T$ is a target with 1 being the k-th element.

Con of Linear Classification–Masking

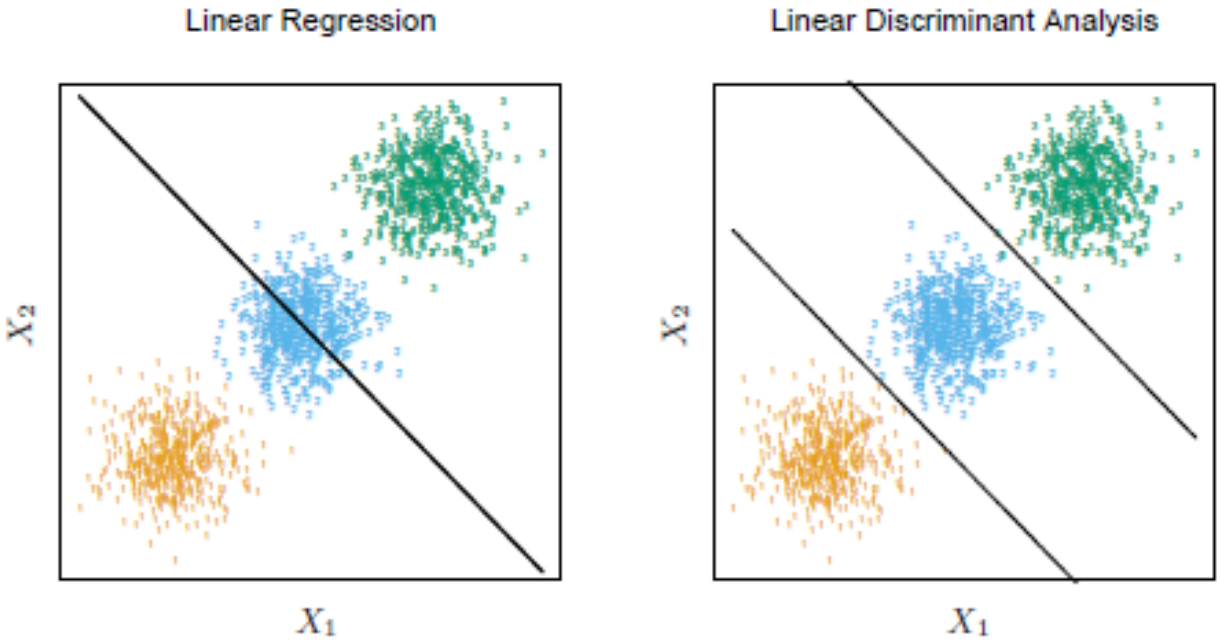

FIGURE 4.2. *The data come from three classes in* $\mathbb{R}^2$ *and are easily separated by linear decision boundaries. The right plot shows the boundaries found by linear discriminant analysis. The left plot shows the boundaries found by linear regression of the indicator response variables. The middle class is completely masked (never dominates).*

Outline

Generative Models

e.g. LDA, QDA

1. Assume sample points come from probability distributions different for each class.
2. For each class C, fit distribution parameters to class C points, giving $P(X|Y = C)$
3. For each class C, estimate $P(Y = C)$
4. Pick class C that maximizes $P(Y = C|X = x)$

Discriminative Models

e.g. Logistic Regression

1. Model $P(Y|X)$ directly.

Gaussian Discriminant Analysis

This is a special generative method in which the class conditional probability distributions are Gaussian: $(\mathbf{X}|Y = k) \sim N(\mu_k, \Sigma_k)$.

Assume that we are given a training set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ of n points. Estimating the prior distribution of class k is $P(k) = \frac{n_k}{n}$, where n_k is the number of training points that belong to class k.

Gaussian Discriminant Analysis

Then we start to estimate the parameters of the conditional distributions with MLE.

$$
\begin{aligned}
\hat{y} &= \arg\max_k p(k|\mathbf{x}) \\
&= \arg\max_k P(k)p(\mathbf{x}|k) \\
&= \arg\max_k \ln(P(k)) + \ln(p(\mathbf{x}|k)) \\
&= \arg\max_k \ln(P(k)) \\
&+ \ln\left(\frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)\right)\right) \\
&= \arg\max_k \ln(P(k)) - \frac{1}{2}\ln(|\Sigma_k|) - \frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)
\end{aligned}
$$

Quadratic Discriminant Analysis(QDA)

The covariance Σ_k of class k has no dependence/relation to that of the other classes. Due to the independence property, we can estimate the true mean and covariance μ_k, Σ_k for each class conditional probability distribution $p(\mathbf{x}|k)$ independently, with the n_k samples in our training data that are classified as class k.

Quadratic Discriminant Analysis(QDA)

$$
\begin{aligned}
\hat{\mu}_k, \hat{\Sigma}_k &= \arg\max_{\mu_k, \Sigma_k} P(X_1, X_2, \cdots, X_{n_k}|k) \\
&= \arg\max_{\mu_k, \Sigma_k} \ln(P(X_1, X_2, \cdots, X_{n_k}|k)) \\
&= \arg\max_{\mu_k, \Sigma_k} \sum_{i=1}^{n_k} \ln(P(X_i|k)) \\
&= \arg\max_{\mu_k, \Sigma_k} \sum_{i=1}^{n_k} -\frac{1}{2}\ln(|\Sigma_k|) - \frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)
\end{aligned}
$$

$$
\frac{\partial f(\mu_k, \Sigma_k)}{\partial \mu_k} = 0 \Rightarrow \hat{\mu}_k = \frac{1}{n_k}\sum_{i:y_i=k} \mathbf{x}_i
$$

$$
\frac{\partial f(\mu_k, \Sigma_k)}{\partial \Sigma_k} = 0 \Rightarrow \hat{\Sigma}_k = \frac{1}{n_k}\sum_{i:y_i=k} (\mathbf{x}_i - \hat{\mu}_{y_i})(\mathbf{x}_i - \hat{\mu}_{y_i})^T
$$

Linear Discriminant Analysis (LDA)

LDA assumes that the class conditional probability distributions are normally distributed with different means μ_k, but LDA is different from QDA in that it requires all of the distributions to share the same covariance matrix Σ.

Linear Discriminant Analysis (LDA)

$$\hat{\mu}_k, \hat{\Sigma} = \arg\max_{\mu_k, \Sigma} \prod_{i:y_i=k} P(X_1, X_2, \cdots, X_{n_k}|k)$$

$$= \arg\max_{\mu_k, \Sigma} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \ln(P(X_i|k))$$

$$= \arg\max_{\mu_k, \Sigma} \sum_{k=1}^{K} \sum_{i=1}^{n_k} -\frac{1}{2}\ln(|\Sigma|) - \frac{1}{2}(\mathbf{x} - \mu_k)^T \Sigma^{-1}(\mathbf{x} - \mu_k)$$

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} \mathbf{x}_i$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\mu}_{y_i})(\mathbf{x}_i - \hat{\mu}_{y_i})^T$$

LDA and QDA

But in the real implement, if we use Σ and Σ_k above, we will get biased estimate. If we want the unbiased estimate, we need to use the pooled variance and sample variance.

$$\hat{\Sigma}_k = \frac{1}{n_k - 1} \sum_{i:y_i=k} (\mathbf{x}_i - \hat{\mu}_{y_i})(\mathbf{x}_i - \hat{\mu}_{y_i})^T$$

$$\hat{\Sigma} = \frac{1}{n - K} \sum_{k=1}^{K} \sum_{i:y_i=k} (\mathbf{x}_i - \hat{\mu}_k)(\mathbf{x}_i - \hat{\mu}_k)^T$$

Sample Variance

In this slide, I will prove why $\hat{\Sigma}_k = \frac{1}{n_k} \sum_{i:y_i=k} (\mathbf{x}_i - \hat{\mu}_{y_i})(\mathbf{x}_i - \hat{\mu}_{y_i})^T$ is biased and why sample variance is unbiased. Let's simplify this problem to one dimension.

- Total Mean: μ
- Total Variance: σ^2
- Sample Mean: $\hat{\mu}$
- Sample Variance: s^2

We want $E(s^2) = \sigma^2$

Sample Variance

If $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$

$$
\begin{aligned}
E(s^2) &= E[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2] \\
&= E[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu + \mu - \hat{\mu})^2] \\
&= E[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{2}{n}\sum_{i=1}^{n}(x_i - \mu)(\hat{\mu} - \mu) + \frac{1}{n}\sum_{i=1}^{n}(\mu - \hat{\mu})^2] \\
&= E[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 - 2(\hat{\mu} - \mu)(\hat{\mu} - \mu) + (\mu - \hat{\mu})^2] \\
&= E[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 - (\mu - \hat{\mu})^2] \\
&= E[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2] - E[(\mu - \hat{\mu})^2] \\
&= \sigma^2 - E[\frac{1}{n}(\mu - x_i)^2] \\
&= \sigma^2 - \frac{1}{n}\sigma^2 < \sigma^2
\end{aligned}
$$

Sample Variance

If $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$

$$
\begin{aligned}
E(s^2) &= E[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2] \\
&= E[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu + \mu - \hat{\mu})^2] \\
&= E[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{2}{n-1}\sum_{i=1}^{n}(x_i - \mu)(\hat{\mu} - \mu) + \frac{1}{n-1}\sum_{i=1}^{n}(\mu - \hat{\mu})^2] \\
&= E[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{2n}{n-1}(\hat{\mu} - \mu)(\hat{\mu} - \mu) + \frac{n}{n-1}(\mu - \hat{\mu})^2] \\
&= E[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{n-1}(\mu - \hat{\mu})^2] \\
&= E[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2] - E[\frac{n}{n-1}(\mu - \hat{\mu})^2] \\
&= E[\frac{n}{n-1}\sigma^2 - \frac{n}{n-1}\frac{\sigma^2}{n}] = \sigma^2
\end{aligned}
$$

LDA and QDA Decision Boundary

The term quadratic in QDA and linear in LDA actually signify the shape of the decision boundary. We will prove this claim using binary (2-class) examples for simplicity (class A and class B). An arbitrary point $\mathbf{x}$ is classified according to the following cases.

$$\hat{y} = \begin{cases} A & p(\mathbf{x}|A) > p(\mathbf{x}|B) \\ B & p(\mathbf{x}|A) < p(\mathbf{x}|B) \\ \text{Either } A \text{ or } B & p(\mathbf{x}|A) = p(\mathbf{x}|B) \end{cases}$$

The decision boundary is the set of all points in x-space that are classified according to the third case.

LDA Decision Boundary

$$
\begin{aligned}
p(\mathbf{x}|A) &= p(\mathbf{x}|B) \\
\ln p(\mathbf{x}|A) &= \ln p(\mathbf{x}|B) \\
\ln(P(A)) - \frac{1}{2}\ln(|\hat{\Sigma}_A|) - \frac{1}{2}(\mathbf{x}-\hat{\mu}_A)^T\hat{\Sigma}_A^{-1}(\mathbf{x}-\hat{\mu}_A) &= \ln(P(B)) - \frac{1}{2}\ln(|\hat{\Sigma}_B|) - \frac{1}{2}(\mathbf{x}-\hat{\mu}_B)^T\hat{\Sigma}_B^{-1}(\mathbf{x}-\hat{\mu}_B) \\
\ln(P(A)) - \frac{1}{2}\ln(|\hat{\Sigma}|) - \frac{1}{2}(\mathbf{x}-\hat{\mu}_A)^T\hat{\Sigma}^{-1}(\mathbf{x}-\hat{\mu}_A) &= \ln(P(B)) - \frac{1}{2}\ln(|\hat{\Sigma}|) - \frac{1}{2}(\mathbf{x}-\hat{\mu}_B)^T\hat{\Sigma}^{-1}(\mathbf{x}-\hat{\mu}_B) \\
\ln(P(A)) - \frac{1}{2}(\mathbf{x}-\hat{\mu}_A)^T\hat{\Sigma}^{-1}(\mathbf{x}-\hat{\mu}_A) &= \ln(P(B)) - \frac{1}{2}(\mathbf{x}-\hat{\mu}_B)^T\hat{\Sigma}^{-1}(\mathbf{x}-\hat{\mu}_B) \\
2\ln(P(A)) - \mathbf{x}^T\hat{\Sigma}^{-1}\mathbf{x} + 2\mathbf{x}^T\hat{\Sigma}^{-1}\hat{\mu}_A - \hat{\mu}_A\hat{\Sigma}^{-1}\hat{\mu}_A &= 2\ln(P(B)) - \mathbf{x}^T\hat{\Sigma}^{-1}\mathbf{x} + 2\mathbf{x}^T\hat{\Sigma}^{-1}\hat{\mu}_B - \hat{\mu}_B\hat{\Sigma}^{-1}\hat{\mu}_B \\
2\ln(P(A)) + 2\mathbf{x}^T\hat{\Sigma}^{-1}\hat{\mu}_A - \hat{\mu}_A\hat{\Sigma}^{-1}\hat{\mu}_A &= 2\ln(P(B)) + 2\mathbf{x}^T\hat{\Sigma}^{-1}\hat{\mu}_B - \hat{\mu}_B\hat{\Sigma}^{-1}\hat{\mu}_B
\end{aligned}
$$

The decision boundary

$$
\mathbf{x}^T\hat{\Sigma}^{-1}(\hat{\mu}_A - \hat{\mu}_B) + \left(\ln(\frac{P(A)}{P(B)}) - \frac{\hat{\mu}_A\hat{\Sigma}^{-1}\hat{\mu}_A - \hat{\mu}_B\hat{\Sigma}^{-1}\hat{\mu}_B}{2}\right) = 0
$$

Binary Logistic Regression

We view each observasions y_i as an independent sample from a Bernoulli distribution $Y_i \sim Bern(p_i)$, (technically we mean $\hat{Y}_i|\mathbf{x}_i, \mathbf{w}$), where p_i is a function of $\mathbf{x}_i$.

We need a model for the dependency of p_i on $\mathbf{x}_i$. We have to enforce that p_i is a transformation of $\mathbf{x}_i$ that results in a number from 0 to 1 (ie. a valid probability). Hence p_i cannot be, say, linear in x_i. One way to do achieve the 0-1 normalization is by using the sigmoid function.

$$p_i = s(\mathbf{w}^T \mathbf{x}_i) = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}_i}}$$

Binary Logistic Regression

Now we can estimate the parameters $\mathbf{w}$ via maximum likelihood.

$$
\begin{aligned}
\hat{\mathbf{w}} &= \arg\max_{\mathbf{w}} P(y_1, \cdots, y_n | \mathbf{x}_1, \cdots, \mathbf{x}_n, \mathbf{w}) \\
&= \arg\max_{\mathbf{w}} \prod_{i=1}^{n} P(y_i | \mathbf{x}_i, \mathbf{w}) \\
&= \arg\max_{\mathbf{w}} \ln[\prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{(1-y_i)}] \\
&= \arg\max_{\mathbf{w}} \sum_{i=1}^{n} y_i \ln p_i + (1 - y_i) \ln(1 - p_i) \\
&= \arg\min_{\mathbf{w}} - \sum_{i=1}^{n} y_i \ln p_i + (1 - y_i) \ln(1 - p_i)
\end{aligned}
$$

Binary Logistic Regression

The logistic regression loss function has no known analytic closed-form solution. Therefore, in order to minimize it, we can use gradient descent, either in batch form or stochastic form. Let's examine the case for batch gradient descent.

Binary Logistic Regression

$$L(\mathbf{w}) = -\sum_{i=1}^{n} y_i \ln p_i + (1 - y_i) \ln(1 - p_i)$$

$$\begin{aligned}
\nabla_{\mathbf{w}} L(\mathbf{w}) &= \nabla_{\mathbf{w}} \left(-\sum_{i=1}^{n} y_i \ln p_i + (1 - y_i) \ln(1 - p_i) \right) \\
&= -\sum_{i=1}^{n} y_i \nabla_{\mathbf{w}} \ln p_i + (1 - y_i) \nabla_{\mathbf{w}} \ln(1 - p_i) \\
&= -\sum_{i=1}^{n} \frac{y_i}{p_i} \nabla_{\mathbf{w}} p_i - \frac{1 - y_i}{1 - p_i} \nabla_{\mathbf{w}} p_i
\end{aligned}$$

Binary Logistic Regression

$$
\begin{aligned}
p_i = s(\mathbf{w}^T x_i) &= \frac{1}{1+e^{-\mathbf{w}^T x_i}} \\
\nabla_z s(z) &= \nabla_z (1+e^{-z})^{-1} \\
&= \frac{e^{-z}}{(1+e^{-z})^2} \\
&= s(z)(1-s(z)) \\
\nabla_{\mathbf{w}} L(\mathbf{w}) &= -\sum_{i=1}^{n} \frac{y_i}{p_i} \nabla_{\mathbf{w}} p_i - \frac{1-y_i}{1-p_i} \nabla_{\mathbf{w}} p_i \\
&= -\sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i} \right) p_i(1-p_i) x_i \\
&= -\sum_{i=1}^{n} (y_i - p_i) x_i
\end{aligned}
$$

Binary Logistic Regression

Hessian of loss function.

$$
\begin{aligned}
H_{kl} &= \frac{\partial^2 L(\mathbf{w})}{\partial \mathbf{w}_k \partial \mathbf{w}_l} \\
&= \frac{\partial}{\partial \mathbf{w}_k} - \sum_{i=1}^{n} (y_i - p_i)\mathbf{x}_{il} \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \mathbf{w}_k} p_i x_{il} \\
&= \sum_{i=1}^{n} p_i(1 - p_i) x_{ik} x_{il} \\
H &= \sum_{i=1}^{n} p_i(1 - p_i) x_i x_i^T \\
&= XWX^T
\end{aligned}
$$

LDA vs Logistic Regression

1. LDA model can always be written as a logistic regression (LR) model.
2. However, the converse is not true. Not every logistic regression model can be written as an LDA model.
3. Thus LDA makes strictly stronger modelling assumptions than LR, and LR is a more general model.
4. When LDA assumptions are correct (class-conditional densities are Gaussian), LDA tends to be better than LR.
5. As data goes to infinity, results more similar.

ESL EX4.1

We want to maximize the between class variance relative to the with in class variance

$$\max_a \frac{a^T \mathbf{B} a}{a^T \mathbf{W} a}$$

where

$$\mathbf{B} = \sum_{k=1}^{K} N_k (\mu_k - \overline{\mu})(\mu_k - \overline{u})^T$$

$$\mathbf{W} = \sum_{k=1}^{K} \sum_{g_i = k} (x_i - \overline{\mu}_k)(x_i - \overline{\mu}_k)^T$$

ESL EX4.1

$$
\begin{aligned}
J(a) &= \frac{a^T \mathbf{B} a}{a^T \mathbf{W} a} \\
\frac{dJ(a)}{da} &= \frac{d}{da}\left(\frac{a^T \mathbf{B} a}{a^T \mathbf{W} a}\right) = 0 \\
&\Rightarrow (a^T \mathbf{W} a) 2\mathbf{B} a - (a^T \mathbf{B} a) 2\mathbf{W} a = 0 \\
&\Rightarrow \mathbf{B} a - \left(\frac{a^T \mathbf{B} a}{a^T \mathbf{W} a}\right) \mathbf{W} a = 0 \\
&\Rightarrow \mathbf{B} a - J(a) \mathbf{W} a = 0 \\
&\Rightarrow J(a) a = \mathbf{W}^{-1} B a
\end{aligned}
$$

Then a will be the eigen vectors of $X = \mathbf{W}^{-1}\mathbf{B}$, $J(a)$ is eigen value of X.