CSE 6363 - Machine Learning

Linear Discriminant Analysis

Alex Dillhoff

University of Texas at Arlington

What is covered?

1. Formulation for binary classification
2. Gaussian Class Conditional Densities
3. Estimating parameters via MLE
4. Example using `scikit-learn`

Discriminant Functions

Linear Discriminant Analysis

Discriminative functions assign each input vector $\mathbf{x}$ to a class depending on whether the output met a particular threshold.

Modeling the conditional probability distribution $p(C_k|\mathbf{x})$ grants us additional benefits while still fulfilling our original classification task.

Linear Discriminant Analysis

Let's begin with a 2 class problem.

To classify this with a generative model, we use the class-conditional densities $p(\mathbf{x}|C_i)$ and class priors $p(C_i)$.

Linear Discriminant Analysis

The posterior probability for C_1 can be written in the form of a sigmoid function:

$$p(C_1|\mathbf{x}) = \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_1)p(C_1) + p(\mathbf{x}|C_2)p(C_2)}$$

Linear Discriminant Analysis

Then multiply the numerator and denominator by

$$\frac{(p(\mathbf{x}|C_1))^{-1}}{(p(\mathbf{x}|C_1))^{-1}}.$$

Linear Discriminant Analysis

This yields

$$\frac{1}{1 + \frac{p(\mathbf{x}|C_2)p(C_2)}{p(\mathbf{x}|C_1)p(C_1)}}.$$

Linear Discriminant Analysis

Noting that $a = \exp(\ln(a))$, we can rewrite further

$$\frac{1}{1 + \exp(-a)},$$

where $a = \ln \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_2)p(C_2)}$.

Linear Discriminant Analysis

Writing this distribution in the form of the sigmoid function is convenient as it is a natural choice for many other classification models.

It also has a very simple derivative which is convenient for models optimized using gradient descent.

Linear Discriminant Analysis

Given certain choices for the class conditional densities, the posterior probabilty distribution will be a linear function of the input features:

$$\ln p(C_k|\mathbf{x};\theta) = \mathbf{w}^T\mathbf{x} + c,$$

Linear Discriminant Analysis

$$\ln p(C_k|\mathbf{x};\theta) = \mathbf{w}^T\mathbf{x} + c,$$

where **w** is a parameter vector based on the parameters of the chosen probability distribution, and c is a constant term that is not dependent on the parameters.

Gaussian Class Conditional Densities

Linear Discriminant Analysis

What do we mean by **"certain choices for the class conditional densities?"**

One convenient choice is to use **Gaussian Class Conditional Densities.**

Linear Discriminant Analysis

Let's assume that our class conditional densities $p(\mathbf{x}|C_k)$ are Gaussian.

We will additionally assume that the covariance matrices between classes are shared.

This will result in linear decision boundaries.

Linear Discriminant Analysis

Since the conditional densities are chosen to be Gaussian, the posterior is given by

$$p(C_k|\mathbf{x};\theta) \propto \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c, \Sigma),$$

where π_k is the prior probability of class k.

We can ignore the normalizing constant since it is not dependent on the class.

Linear Discriminant Analysis

Since the conditional densities are chosen to be Gaussian, the posterior is given by

$$p(C_k|\mathbf{x};\theta) \propto \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c, \Sigma),$$

where π_k is the prior probability of class k.

We can ignore the normalizing constant since it is not dependent on the class.

Linear Discriminant Analysis

The class conditional density function for class k is given by

$$p(\mathbf{x}|C_k;\theta) = \frac{1}{2\pi^{D/2}}\frac{1}{|\Sigma|^{1/2}}\exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\Big).$$

Linear Discriminant Analysis

Let's go back to the simple case of two classes and define $a = \ln \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_2)p(C_2)}$.

First, we rewrite a:

$$a = \ln p(\mathbf{x}|C_1) - \ln p(\mathbf{x}|C_2) + \ln \frac{p(C_1)}{p(C_2)}.$$

Linear Discriminant Analysis

The log of the class conditional density for a Gaussian is

$$\ln p(\mathbf{x}|C_k;\boldsymbol{\mu}_k,\Sigma) = -\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_k).$$

Linear Discriminant Analysis

To simplify the above result, we will group the terms that are not dependent on the class parameters since they are consant:

$$\ln p(\mathbf{x}|C_k; \boldsymbol{\mu}_k, \Sigma) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) + c.$$

Linear Discriminant Analysis

Observing that this quantity takes on a quadratic form, we can rewrite the above as

$$\ln p(\mathbf{x}|C_k; \boldsymbol{\mu}_k, \Sigma) = -\frac{1}{2}\boldsymbol{\mu}_k \Sigma^{-1} \boldsymbol{\mu}_k + \mathbf{x}^T \Sigma^{-1} \boldsymbol{\mu}_k - \frac{1}{2}\mathbf{x}^T \Sigma^{-1} \mathbf{x} + c.$$

Linear Discriminant Analysis

Using this, we complete the definition of a:

$$
\begin{aligned}
a &= \ln p(\mathbf{x}|C_1) - \ln p(\mathbf{x}|C_2) + \ln \frac{p(C_1)}{p(C_2)} \\
&= -\frac{1}{2}\boldsymbol{\mu}_1\Sigma^{-1}\boldsymbol{\mu}_1 + \mathbf{x}^T\Sigma^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2\Sigma^{-1}\boldsymbol{\mu}_2 - \mathbf{x}^T\Sigma^{-1}\boldsymbol{\mu}_2 + \ln \frac{p(C_1)}{p(C_2)} \\
&= \mathbf{x}^T(\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)) - \frac{1}{2}\boldsymbol{\mu}_1\Sigma^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2\Sigma^{-1}\boldsymbol{\mu}_2 + \ln \frac{p(C_1)}{p(C_2)} \\
&= (\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^T\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_1\Sigma^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2\Sigma^{-1}\boldsymbol{\mu}_2 + \ln \frac{p(C_1)}{p(C_2)}.
\end{aligned}
$$

Linear Discriminant Analysis

Finally, we define

$$\mathbf{w} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

and

$$w_0 = -\frac{1}{2}\boldsymbol{\mu}_1\Sigma^{-1}\boldsymbol{\mu}_1 - \frac{1}{2}\boldsymbol{\mu}_2\Sigma^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(C_1)}{p(C_2)}.$$

Linear Discriminant Analysis

Thus, our posterior takes on the form

$$p(C_1|\mathbf{x};\theta) = \sigma(\mathbf{w}^T\mathbf{x} + w_0).$$

Multiple Classes

Multiple Classes

What if we have more than 2 classes?

Recall that a **generative classifier** is modeled as

$$p(C_k|\mathbf{x};\boldsymbol{\theta}) = \frac{p(C_k|\boldsymbol{\theta})p(\mathbf{x}|C_k,\boldsymbol{\theta})}{\sum_{k'} p(C_{k'}|\boldsymbol{\theta})p(\mathbf{x}|C_{k'},\boldsymbol{\theta})}.$$

Multiple Classes

As stated previously, $\boldsymbol{\pi}_k = p(C_k|\boldsymbol{\theta})$ and $p(\mathbf{x}|C_k, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c, \Sigma)$.

For LDA, the covariance matrices are shared across all classes.

Multiple Classes

This permits a simplification of the class posterior distribution $p(C_k|\mathbf{x};\boldsymbol{\theta})$:

$$\begin{aligned} p(C_k|\mathbf{x};\boldsymbol{\theta}) &\propto \boldsymbol{\pi}_k \exp\left(\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_k\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k\right) \\ &= \exp\left(\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_k\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log\boldsymbol{\pi}_k\right)\exp\left(-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right). \end{aligned}$$

Multiple Classes

The term $\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}\right)$ is placed aside since it is not dependent on the class k.

When divided by the sum per the definition of $p(C_k|\mathbf{x};\boldsymbol{\theta})$, it will equal to 1.

Multiple Classes

Under this formulation, we let

$$\mathbf{w}_k = \Sigma^{-1}\boldsymbol{\mu}_k$$
$$\mathbf{b}_k = -\frac{1}{2}\boldsymbol{\mu}_k^T\Sigma^{-1}\boldsymbol{\mu}_k + \log \boldsymbol{\pi}_k.$$

Multiple Classes

This lets us express $p(C_k|\mathbf{x};\boldsymbol{\theta})$ as the **softmax** function:

$$p(C_k|\mathbf{x};\boldsymbol{\theta}) = \frac{\exp(\mathbf{w}_k^T\mathbf{x}+\mathbf{b}_k)}{\sum_{k'}\exp(\mathbf{w}_{k'}^T\mathbf{x}+\mathbf{b}_{k'})}.$$

Decision Boundaries

Decision Boundaries

Classifications can be made by choosing the class with the highest posterior probability.

Geometrically, this decision boundary has a direct connection to logistic regression.

The decision boundary is the set of points where the posterior probability of two classes is equal.

This is the set of points where the linear discriminant function is equal to 0.

Decision Boundaries

In the previous section, the derivation for the posterior probability of class C_k was written in the form of the softmax function

$$p(C_k|\mathbf{x};\theta) = \frac{\exp(\mathbf{w}_k^T\mathbf{x} + \mathbf{b}_k)}{\sum_{k'}\exp(\mathbf{w}_{k'}^T\mathbf{x} + \mathbf{b}_{k'})}.$$

Decision Boundaries

In the binary case, the posterior for class 1 is given by

$$
\begin{aligned}
p(C_1|\mathbf{x};\theta) &= \frac{\exp(\mathbf{w}_1^T\mathbf{x}+\mathbf{b}_1)}{\exp(\mathbf{w}_1^T\mathbf{x}+\mathbf{b}_1)+\exp(\mathbf{w}_2^T\mathbf{x}+\mathbf{b}_2)} \\
&= \frac{1}{1+\exp((\mathbf{w}_1-\mathbf{w}_2)^T\mathbf{x}+(\mathbf{b}_1-\mathbf{b}_2))} \\
&= \sigma((\mathbf{w}_1-\mathbf{w}_2)^T\mathbf{x}+(\mathbf{b}_1-\mathbf{b}_2)).
\end{aligned}
$$

Decision Boundaries

Using the previous definition of $\mathbf{b}_k$, we can rewrite $\mathbf{b}_1 - \mathbf{b}_2$ as

$$\begin{aligned}
\mathbf{b}_1 - \mathbf{b}_2 &= -\frac{1}{2}\mu_1^T \mathbf{\Sigma}^{-1}\mu_1 + \log \pi_1 + \frac{1}{2}\mu_2^T \mathbf{\Sigma}^{-1}\mu_2 - \log \pi_2 \\
&= -\frac{1}{2}(\mu_1 - \mu_2)^T \mathbf{\Sigma}^{-1}(\mu_1 + \mu_2) + \log \frac{\pi_1}{\pi_2}
\end{aligned}$$

Decision Boundaries

This can be used to define a new weight vector $\mathbf{w}'$ and a point directly between the two class means $\mathbf{x}_0$:

$$\mathbf{w}' = \boldsymbol{\Sigma}^{-1}(\mu_1 - \mu_2)$$
$$\mathbf{x}_0 = \frac{1}{2}(\mu_1 + \mu_2) - (\mu_1 - \mu_2)\frac{\log \frac{\pi_1}{\pi_2}}{(\mu_1 - \mu_2)^T \boldsymbol{\Sigma}^{-1}(\mu_1 - \mu_2)}.$$

Decision Boundaries

With these new terms defined, we have that $\mathbf{w}'^T\mathbf{x}_0 = -(\mathbf{b}_1 - \mathbf{b}_2)$ and the posterior probability for class 1 can be written in the form of binary logistic regression:

$$p(C_1|\mathbf{x};\theta) = \sigma(\mathbf{w}'^T(\mathbf{x} - \mathbf{x}_0)).$$

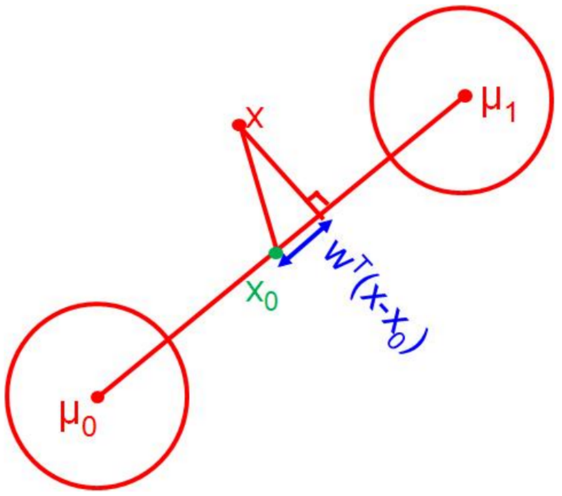
Decision Boundaries

- The middle point between the two class means $\mathbf{x}_0$ is the point where the posterior probability of class 1 is 0.5.
- This is the decision boundary between the two classes.
- If $\mathbf{w}'^T\mathbf{x} > \mathbf{w}'^T\mathbf{x}_0$, then the posterior probability of class 1 is greater than 0.5 and the input vector $\mathbf{x}$ is classified as class 1.

Decision Boundaries

- The split between the class priors controls the location of the decision boundary.
- If the class priors are equal, then the decision boundary is the point directly between the two class means.
- If the class priors are not equal, then the decision boundary is shifted towards the class with the higher prior.

Decision Boundaries



Decision boundary between two classes (Murphy, 2022).

Maximum Likelihood Estimation

Maximum Likelihood Estimation

Given this formulation using Gaussian densities, we can estimate the parameters of the model via **maximum likelihood estimation**.

Assuming K classes with Gaussian class conditional densities, the likelihood function is

$$p(\mathbf{X}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathcal{M}(y_i|\boldsymbol{\pi}) \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{\mathbb{1}(y_i=k)}.$$

Maximum Likelihood Estimation

Taking the log of this function yields

$$\ln p(\mathbf{X}|\theta) = \Big[\sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{1}(y_i = k)\ln \pi_k\Big] + \sum_{k=1}^{K}\Big[\sum_{i:y_i=c} \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\Big].$$

Class Prior

For multinomial distributions, the class prior parameter estimation $\hat{\pi}_k$ is easily calculated by counting the number of samples belonging to class k and dividing it by the total number of samples.

$$\hat{\pi}_k = \frac{n_k}{n}$$

Class Prior

The parameter estimates are

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

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

Quadratic Discriminant Analysis

Quadratic Discriminant Analysis

Linear Discriminant Analysis is a special case of Quadratic Discriminant Analysis (QDA) where the covariance matrices are shared across all classes.

Assuming each class conditional density is Gaussian, the posterior probability is given by

$$p(C_k|\mathbf{x};\theta) \propto \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k).$$

Quadratic Discriminant Analysis

Taking the log of this function yields

$$\ln p(C_k|\mathbf{x};\theta) = \ln \pi_k - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k) + c.$$

Quadratic Discriminant Analysis

With LDA, the term $\frac{1}{2} \ln |\Sigma_k|$ is constant across all classes, so we treat it as another constant.

Since QDA considers a different covariance matrix for each class, we must keep this term in the equation.

Quadratic Discriminant Analysis

In the more general case of QDA, the decision boundary is quadratic, leading to a quadratic discriminant function.

As shown in the previously, the posterior probability function for LDA is linear in $\mathbf{x}$, which leads to a linear discriminant function.

Summary

Summary

- LDA is a generative classifier that assumes Gaussian class conditional densities.
- LDA assumes that the covariance matrices are shared across all classes.
- LDA can be extended to multiple classes.
- LDA can be estimated via maximum likelihood estimation.

Summary

Given some data $\mathbf{X}$ and labels $\mathbf{y}$, we can estimate the parameters of the model via maximum likelihood estimation.

$$\hat{\pi}_k = \frac{n_k}{n}$$

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

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

Summary

Using these estimates, we can compute the weights and biases for our linear discriminant functions.

$$\mathbf{w}_k = \Sigma^{-1}\boldsymbol{\mu}_k$$
$$\mathbf{b}_k = -\frac{1}{2}\boldsymbol{\mu}_k^T\Sigma^{-1}\boldsymbol{\mu}_k + \log \boldsymbol{\pi}_k.$$

Summary

Given these weights and biases, we can compute the class posterior probabilities using the softmax function.

$$p(C_k|\mathbf{x};\boldsymbol{\theta}) = \frac{\exp(\mathbf{w}_k^T\mathbf{x} + \mathbf{b}_k)}{\sum_{k'}\exp(\mathbf{w}_{k'}^T\mathbf{x} + \mathbf{b}_{k'})}$$

Summary

The class with the highest posterior probability is the class that is predicted.

$$\hat{y} = \text{argmax}_k p(C_k|\mathbf{x}; \boldsymbol{\theta})$$