Lecture 10

Linear Models for Classification – Part 2
CS542—Contents 10

- Classification
  - Least squares discrimination
  - Fisher’s Linear Discriminant Analysis
  - Probabilistic Generative Models

- Handout:
  - Class Notes

- Reading Assignment for Next Class
  - Bishop Ch.4
Learning Linear Discriminant Functions

- Decision Regions must be simply connected (no holes) and convex!!!
Why not use least squares regression for multi-class classification?

\[ y_i = w_i^T \tilde{x} + w_{i,0} \]

\[ y(x) = Wx \]

\[ t_i = [0,0,\ldots,0,1,0,\ldots,0]^T \]

\[ W = \left( X^T X \right)^{-1} XT \]

However, least squares classification is not robust!
Least Squares Discrimination

- Another example how things can go wrong

The problem is related that in classification, a Gaussian error distribution is hardly the correct cost function.
Fisher’s Linear Discriminant Analysis

- View discrimination in terms of dimensionality reduction:
  - Project data and try to discriminate in projection
    \[ y = \mathbf{w}^T \mathbf{x} \]
  - If \( y \geq -w_0 \) it is class 1, otherwise class 2
  - But this kind of projection may not lead to good discrimination
Fisher’s Linear Discriminant Analysis

- Deriving a better way of classification:
  - Assume two classes, with \( N_1 \) and \( N_2 \) data points per class.
  - Define mean vectors:
    \[
    \mathbf{m}_1 = \frac{1}{N_1} \sum_{n \in C_1} \mathbf{x}_n \quad \mathbf{m}_2 = \frac{1}{N_2} \sum_{n \in C_2} \mathbf{x}_n
    \]
  - Measure separation of classes from separation of means
    \[
    m_2 - m_1 = \mathbf{w}^T (\mathbf{m}_2 - \mathbf{m}_1)
    \]
    - i.e., choose \( \mathbf{w} \) to maximize this separation, while constraining \( \mathbf{w} \) to unit length
    \[
    \|\mathbf{w}\| = 1
    \]
Fisher’s Linear Discriminant Analysis

- Thus, we need to optimize:

\[ J(w) = w^T (m_2 - m_1) + \lambda (w^T w - 1) \]

- Resulting in:

\[ w = \frac{m_2 - m_1}{\|m_2 - m_1\|} \]

- Problem: the variance in the project may lead to significant overlap of the distributions:
Fisher’s Linear Discriminant Analysis

- Thus, we need to optimize taking the variance in the projection into account:
  - The idea of Fisher was to maximize the separation of the means while simultaneously minimizing the variance within each class:

$$S_k^2 = \sum_{n \in C_k} (y_n - m_k)^2$$

- Thus, the new cost function is:

$$J(w) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} + \lambda (w^T w - 1)$$

$$= \frac{w^T S_B w}{w^T S_w w} + \lambda (w^T w - 1)$$

where

$$S_B = (m_2 - m_1)(m_2 - m_1)^T := \text{between class covariance}$$

$$S_w = \sum_{n \in C_1} (x_n - m_1)(x_n - m_1)^T + \sum_{n \in C_2} (x_n - m_2)(x_n - m_2)$$
Fisher’s Linear Discriminant Analysis

- Result of optimization:

\[ w \propto S_W^{-1}(m_2 - m_1) \]
Probabilistic Generative Models

- The generative view is contrast to the discriminative view, as it tries to formulate the cause of the generating process, and then perform inference.
- Back to Bayes’ Rule of classification:

\[
P(C_1 | x) = \frac{P(x | C_1)P(C_1)}{P(x | C_1)P(C_1) + P(x | C_2)P(C_2)}
\]

\[
= \frac{1}{1 + \frac{P(x | C_2)P(C_2)}{P(x | C_1)P(C_1)}}
\]

\[
= \frac{1}{1 + \exp\left(-\ln\frac{P(x | C_1)P(C_1)}{P(x | C_2)P(C_2)}\right)}
\]

\[
= \frac{1}{1 + \exp(-a)}
\]

This is the famous logistic function!
Probabilistic Generative Models

- For a multi-class problem:

\[
P(C_k \mid x) = \frac{P(x \mid C_k)P(C_k)}{\sum_j P(x \mid C_j)P(C_j)}
\]

This is the normalized exponential or softmax function.
Probabilistic Generative Models

- The interesting case is when the logistic function is parameterized as
  \[
p(C_1 | x) = \sigma(w^T x + w_0) = \frac{1}{1 + \exp(-w^T x + w_0)}
  \]

- Case 1: Class conditional probabilities are Gaussian
  \[
p(x | C_k) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu_k)^T \Sigma^{-1} (x - \mu_k)\right)
  \]

- Bayes Rule demonstrates that the posterior can be parameterized as:
  \[
w = \Sigma^{-1}(\mu_1 - \mu_2)
  \]
  \[
w_0 = -\frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_2^T \Sigma^{-1} \mu_2 + \ln \frac{p(C_1)}{p(C_2)}
  \]
Probabilistic Generative Models

- Thus, the logistic function with $w^T x + w_0$ as argument represents a Bayes optimal decision boundary (assuming that class conditionals are Gaussian with common covariance)
Probabilistic Generative Models

- For the multi-class case we have

\[
P(C_k \mid x) = \frac{\exp(a_k)}{\sum_j \exp(a_j)} \quad \text{where} \quad a_k = \mathbf{w}_k^T \mathbf{x} + w_{k0}
\]

\[
\mathbf{w}_k = \Sigma^{-1} \mu_k
\]

\[
w_{k0} = -\frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \ln p(C_k)
\]
Probabilistic Generative Models

- Note that non-shared covariance matrices will require quadratic decision boundaries for Bayes optimality.
Probabilistic Discriminative Models

- Parameterize the discriminative function and estimate the parameters by maximum likelihood estimation.
- Using general linear model is useful:

\[ \mathbf{w}^T \phi(\mathbf{x}) \]
Probabilistic Discriminative Models

- Logistic Regression

\[ p(C_1 \mid x) = \sigma(w^T \phi(x)) \]

- Maximum Likelihood approach:

\[
p(t \mid w) = \prod_{n=1}^{N} y_n^{t_n} (1 - y_n)^{1-t_n}
\]

\[
\ln p(t \mid w) = \sum (t_n \ln y_n + (1 - t_n) \ln(1 - y_n))
\]

- Solution:
  - Perform gradient descent in \( w \) (use recursive least squares)
  - Laplace approximation
  - Variational approximation

Cross Entropy
Summary

- Linear discriminant analysis is a very useful tool for classification, but needs to be done “right”
- Linear discriminant analysis with a logistic function is motivated from Bayes’ optimality
- Probabilistic discriminant analysis focuses on using Max. Likelihood approaches, with generalized linear models.