Linear Discrimination

CptS 570 – Machine Learning
School of EECS
Washington State University
Linear Discrimination

- Assume instances of classes are linearly separable
- Estimate parameters of linear discriminant

\[
\text{If } (x_2 - 2x_1 - 1) > 0 \quad \text{Then } + \\
\text{Else } -
\]
Discriminant–based vs. Likelihood–based Classification

- Classification (K classes)
  \[ \text{choose } C_i \text{ if } g_i(x) = \max_{j=1}^{K} g_j(x) \]

- Likelihood–based classification
  - Estimate priors \( P(C_i) \) and likelihoods \( p(x|C_i) \)
  - Define \( g_i(x) \) in terms of the posteriors
    \[ g_i(x) = \log P(C_i | x) \]

- Requires knowledge of types of densities
Discriminant–based vs. Likelihood–based Classification

- Discriminant–based classification
  - Learn model of boundaries between classes, instead of densities of bounded regions
    \[ g_i(x | \Phi_i) \]
  - Where \( \Phi_i \) are model parameters of the boundary
Linear Discriminant

- Simple
- Requires only $O(d)$ space to store and $O(d)$ time for classification
- Weight $w_i$ indicates importance of feature $x_i$
- Try linear model before trying more complicated model

$$g_i(x \mid w_i, w_{i0}) = w_i^T x + w_{i0} = \sum_{j=1}^{d} w_{ij} x_j + w_{i0}$$
Generalized Linear Model

- If linear model insufficient, could consider higher-order discriminants
  - More time and space requirements
  - May overfit (bias/variance dilemma)
- Alternative: linear model of non-linear features

\[ z_1 = x_1, \ z_2 = x_2, \ z_3 = x_1^2, \ z_4 = x_2^2, \ z_5 = x_1 x_2 \]

- New features \( z_j \) comprise basis functions \( \phi_j(x) \)

\[ g_i(x) = \sum_{j=1}^{k} w_{ij} \phi_j(x) \]
Two-Class Case

- Weight vector $w$ defines a hyperplane dividing the instance space into two regions

$$g(x) = w^T x + w_0$$

Choose $\begin{cases} C_1 & \text{if } g(x) > 0 \\ C_2 & \text{otherwise} \end{cases}$
Multi-Class Case

- $K > 2$ classes
- Learn one hyperplane for each class separating it from all other classes

$$g_i(x | w_i, w_{i0}) = w_i^T x + w_{i0}$$

Choose $C_i$ if

$$g_i(x) = \max_{j=1}^{K} g_j(x)$$
Multi-Class Case: Pairwise Separation

- Learn one hyperplane for each pair of classes
- \(K(K-1)/2\) hyperplanes

\[
g_{ij}(x \mid w_{ij}, w_{ij0}) = w_{ij}^T x + w_{ij0} = 0
\]

\[
g_{ij}(x) = \begin{cases} 
> 0 & \text{if } x \in C_i \\
\leq 0 & \text{if } x \in C_j \\
don't care & \text{otherwise}
\end{cases}
\]

Choose \(C_i\) if

\(\forall j \neq i, g_{ij}(x) > 0\), or

maximizes \(g_i(x) = \sum_{j \neq i} g_{ij}(x)\)
Parametric Discrimination

- If \( p(x \mid C_i) \sim N(\mu_i, \Sigma) \), then
  \[
  g_i(x \mid w_i, w_{i0}) = w_i^T x + w_{i0}
  \]
  \[
  w_i = \Sigma^{-1} \mu_i \quad w_{i0} = -\frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \log P(C_i)
  \]
  where \( \mu_i, \Sigma, \) and \( P(C_i) \) can be estimated from the data

- Two-class case:
  \[
  y \equiv P(C_1 \mid x) \quad \text{and} \quad P(C_2 \mid x) = 1 - y
  \]
  \[
  \text{choose } C_1 \text{ if } \begin{cases}
    y > 0.5 \\
    y/(1-y) > 1 \\
    \log[y/(1-y)] > 0
  \end{cases}
  \]
  Odds of \( y \)

  and \( C_2 \) otherwise

  Log odds of \( y \), or “logit”
For two normal classes with common covariance matrix, “logit” is linear

\[
\text{logit}(P(C_1 \mid x)) = \log \frac{P(C_1 \mid x)}{1 - P(C_1 \mid x)} = \log \frac{P(C_1 \mid x)}{P(C_2 \mid x)} = \log \frac{p(x \mid C_1)}{p(x \mid C_2)} + \log \frac{P(C_1)}{P(C_2)} = \log \frac{(2\pi)^{-d/2} |\Sigma|^{-1/2} \exp \left[ -\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1) \right]}{(2\pi)^{-d/2} |\Sigma|^{-1/2} \exp \left[ -\frac{1}{2}(x - \mu_2)^T \Sigma^{-1}(x - \mu_2) \right]} + \log \frac{P(C_1)}{P(C_2)} = w^T x + w_0
\]

where \( w = \Sigma^{-1}(\mu_1 - \mu_2) \) \( w_0 = -\frac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2) \)
Parametric Discrimination

- Inverse of “logit” is the “logistic” or “sigmoid” function

\[ P(C_1 | x) = \text{sigmoid}(w^T x + w_0) = \frac{1}{1 + \exp[-(w^T x + w_0)]} \]

- \( w \) and \( w_0 \) can be estimated from the data (see previous slide)
- Sigmoid transforms discriminant value to a posterior probability
Logistic (Sigmoid) function

- Can either
  - Calculate $g(x) = w^T x + w_0$, choose $C_1$ if $g(x) > 0$
  - Calculate $y = \text{sigmoid}(w^T x + w_0)$, choose $C_1$ if $y > 0.5$
Gradient Descent

- Discriminant-based classification seeks $w$ minimizing error $E(w|X)$ on training set $X$

$$w^* = \arg\min_w E(w \mid X)$$

- Use iterative optimization method gradient descent to find $w$

- When $E(w)$ is differentiable, compute gradient vector

$$\nabla_w E = \left[ \frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \ldots, \frac{\partial E}{\partial w_d} \right]^T$$
Gradient Descent

- Start with random $w$
- Update $w$ in the opposite direction of the gradient vector
  \[ \Delta w_i = -\eta \frac{\partial E}{\partial w_i}, \forall i \]
  \[ w_i = w_i + \Delta w_i \]
- Distance of update determined by step size (or learning factor) $\eta$
- Continue update until gradient is zero
  - May be a local minimum
Gradient Descent

\[ \Delta w_i = -\eta \frac{\partial E}{\partial w_i}, \forall i \]

\[ w_i = w_i + \Delta w_i \]
Logistic Discrimination

- Derive a differentiable error function
- Two-class: Assume log likelihood ratio is linear

\[
\log \frac{p(x | C_1)}{p(x | C_2)} = w^T x + w_0^o
\]

\[
\text{logit}(P(C_1 | x)) = \log \frac{P(C_1 | x)}{1 - P(C_1 | x)} = \log \frac{p(x | C_1)}{p(x | C_2)} + \log \frac{P(C_1)}{P(C_2)} = w^T x + w_0
\]

where \( w_0 = w_0^o + \log \frac{P(C_1)}{P(C_2)} \)

\[
y = \hat{P}(C_1 | x) = \frac{1}{1 + \exp[-(w^T x + w_0)]}
\]
Logistic Discrimination

- Two-class case ($r^t = 0$ or $1$):

$$\mathcal{X} = \{x^t, r^t\}_t, \quad r^t | x^t \sim \text{Bernoulli} (y^t)$$

$$y = P(C_1 | x) = \frac{1}{1 + \exp[-(w^T x + w_0)]}$$

$$l(w, w_0 | \mathcal{X}) = \prod_t (y^t)^{r^t} (1 - y^t)^{1-r^t}$$

$$E = -\log l$$

$$E(w, w_0 | \mathcal{X}) = -\sum_t r^t \log y^t + (1 - r^t) \log (1 - y^t)$$
Logistic Discrimination

- Training in 2-class case: Gradient descent

\[ E(w, w_0 \mid X) = -\sum_t r^t \log y^t + (1 - r^t) \log (1 - y^t) \]

If \( y = \text{sigmoid}(a) \) \( \frac{dy}{da} = y(1 - y) \)

\[ \Delta w_j = -\eta \frac{\partial E}{\partial w_j} = \eta \sum_t \left( \frac{r^t}{y^t} - \frac{1 - r^t}{1 - y^t} \right) y^t (1 - y^t) x_j^t \]

\[ = \eta \sum_t (r^t - y^t) x_j^t, \quad j = 1, \ldots, d \]

\[ \Delta w_0 = -\eta \frac{\partial E}{\partial w_0} = \eta \sum_t (r^t - y^t) \]
Logistic Discrimination

- Training in 2-class case: Gradient descent
- Initial weights random and near zero
  - Avoids overfitting due to using irrelevant features
- $x^t_0 = 1$

For $j = 0, \ldots, d$

$w_j \leftarrow \text{rand}(-0.01, 0.01)$

Repeat

For $j = 0, \ldots, d$

$\Delta w_j \leftarrow 0$

For $t = 1, \ldots, N$

$o \leftarrow 0$

For $j = 0, \ldots, d$

$o \leftarrow o + w_j x_j^t$

$y \leftarrow \text{sigmoid}(o)$

$\Delta w_j \leftarrow \Delta w_j + (r^t - y)x_j^t$

For $j = 0, \ldots, d$

$w_j \leftarrow w_j + \eta \Delta w_j$

Until convergence
Logistic Discrimination: Example

- Evolution of line and sigmoid at 10, 100, and 1000 iterations
Logistic Discrimination: Multiple Classes

\[ X = \{ x^t, r^t \} \quad r^t | x^t \sim \text{Mult}_K(1, y^t) \quad y^t_i \equiv P(C_i | x^t) \]

\[
\log \frac{p(x | C_i)}{p(x | C_K)} = w_i^T x + w_{i0}^o \quad \text{Let } C_K \text{ be reference class}
\]

\[
y = \hat{P}(C_i | x) = \frac{\exp[w_i^T x + w_{i0}]}{\sum_{j=1}^K \exp[w_j^T x + w_{j0}]} , i = 1, \ldots, K \quad \text{Softmax}
\]

\[
l(\{w_i, w_{i0}\} | X) = \prod_{t} \prod_{i} (y_i^t)^{r_i^t} \]

\[
E(\{w_i, w_{i0}\} | X) = -\sum_{t} r_i^t \log y_i^t
\]

\[
\Delta w_j = \eta \sum_{t} (r_j^t - y_j^t) x^t \quad \Delta w_{j0} = \eta \sum_{t} (r_j^t - y_j^t)
\]
Logistic Discrimination: K Classes

For $i = 1, \ldots, K$, For $j = 0, \ldots, d$, $w_{ij} \leftarrow \text{rand}(-0.01, 0.01)$

Repeat

For $i = 1, \ldots, K$, For $j = 0, \ldots, d$, $\Delta w_{ij} \leftarrow 0$

For $t = 1, \ldots, N$

For $i = 1, \ldots, K$

\[ o_i \leftarrow 0 \]

For $j = 0, \ldots, d$

\[ o_i \leftarrow o_i + w_{ij}x_j^t \]

For $i = 1, \ldots, K$

\[ y_i \leftarrow \exp(o_i) / \sum_k \exp(o_k) \]

For $i = 1, \ldots, K$

For $j = 0, \ldots, d$

\[ \Delta w_{ij} \leftarrow \Delta w_{ij} + (r_i^t - y_i)x_j^t \]

For $i = 1, \ldots, K$

For $j = 0, \ldots, d$

\[ w_{ij} \leftarrow w_{ij} + \eta \Delta w_{ij} \]

Until convergence
Example

[Diagram showing a scatter plot with two lines and a 3D surface graph]

[CptS 570 – Machine Learning]
Example: WEKA SimpleLogistic on Iris

--- Classifier model (full training set) ---

SimpleLogistic:

Class 0:
29.99 +
[petallength] * -9.96 +
[petalwidth] * -5.71

Class 1:
-6.15 +
[sepalwidth] * 1.67 +
[sepalwidth] * 0.82 +
[petallength] * -0.74 +
[petalwidth] * -1.28

Class 2:
-34.94 +
[sepalwidth] * -0.4 +
[sepalwidth] * -3.76 +
[petallength] * 6.27 +
[petalwidth] * 10.89

98.67% accuracy
(2 mistakes)
Summary: Linear Discrimination

- Assume classes can be separated by a hyperplane
- Even if not linearly-separable, can still find a good linear discriminant
- Gradient descent used to learn weights defining hyperplanes
- Logistic discrimination