Overview

- The basic linear model
- Geometry of linear discrimination
  - Two classes
  - Multiple classes
- Logistic Discrimination
- Generalizing the linear model with basis functions
Likelihood- vs. Discriminant-based Classification

- **Likelihood-based:** Assume a model for $p(x|C_i)$, use Bayes’ rule to calculate $P(C_i|x)$
  
  $$g_i(x) = \log P(C_i|x)$$

- **Discriminant-based:** Assume a model for $g_i(x|\Phi_i)$; no density estimation

- Estimating the boundaries is enough; no need to accurately estimate the densities inside the boundaries
Linear Discriminant

• Linear discriminant:

\[ g_i(x \mid w_i, w_{i0}) = w_i^T x + w_{i0} = \sum_{j=1}^{d} w_{ij} x_j + w_{i0} \]

• Advantages:
  • Simple: \( O(d) \) space/computation
  • Knowledge extraction: Weighted sum of attributes; positive/negative weights, magnitudes (credit scoring)
  • Optimal when \( p(x \mid C_i) \) are Gaussian with shared cov matrix; useful when classes are (almost) linearly separable
Generalized Linear Model

- Quadratic discriminant:

\[ g_i(x \mid W_i, w_i, w_{i0}) = x^T W_i x + w_i^T x + w_{i0} \]

- Higher-order (product) terms:

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

Map from \( x \) to \( z \) using nonlinear basis functions and use a linear discriminant in \( z \)-space

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

Key idea (basis functions)
Two Classes

\[ g(x) = g_1(x) - g_2(x) \]

\[ = (w_1^T x + w_{10}) - (w_2^T x + w_{20}) \]

\[ = (w_1 - w_2)^T x + (w_{10} - w_{20}) \]

\[ = w^T x + w_0 \]

choose \( \begin{cases} C_1 & \text{if } g(x) > 0 \\ C_2 & \text{otherwise} \end{cases} \)
Geometry
K Classes, $K > 2$

For problems where each class is linearly separable from all other classes

Choose $C_i$ if

$$g_i(x) = \max_{j=1}^{\kappa} g_j(x)$$

Choose the class whose hyperplane is most distant from the point
Pairwise Separation

For problems where classes are pairwise linearly separable

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

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

Testing:

choose \( C_i \) if \( \forall j \neq i, g_{ij}(x) > 0 \)
From Discriminants to Posteriors

When \( p(x \mid C_i) \sim N(\mu_i, \Sigma) \)

\[
  g_i(x \mid w_i, w_{i0}) = w_i^T x + w_{i0}
\]

\[
  w_i = \Sigma^{-1} \mu_i
\]

\[
  w_{i0} = -\frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \log P(C_i)
\]

\[
  y = P(C_1 \mid x) \text{ and } P(C_2 \mid x) = 1 - y
\]

\[
  \begin{cases}
    y > 0.5 \\
    y/(1-y) > 1 \\
    \log[y/(1-y)] > 0
  \end{cases}
\]

choose \( C_1 \) if \( y > 0.5 \) and \( C_2 \) otherwise

\[ \text{←logit transformation} \]
logit\left( P(C_1 | x) \right) = \log \frac{P(C_1 | x)}{1 - P(C_1 | x)} = \log \frac{P(C_1 | x)}{P(C_2 | x)}

= \log \frac{p(x | C_1)}{p(x | C_2)} + \log \frac{P(C_1)}{P(C_2)}

= \log \frac{(2\pi)^{-d/2} \Sigma^{-1/2}}{(2\pi)^{-d/2} \Sigma^{-1/2}} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu_1}{\Sigma^{-1}} \right)^T \Sigma^{-1} \left( \frac{x - \mu_1}{\Sigma^{-1}} \right) \right] + \log \frac{P(C_1)}{P(C_2)}

= w^T x + w_0

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

The inverse of logit

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

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

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

- $E(w | X)$ is error with parameters $w$ on sample $X$
- $w^* = \arg \min_w E(w | X)$

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

- Gradient-descent:
  Starts from random $w$ and updates $w$ iteratively in the negative direction of gradient
Gradient-Descent

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

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

- Two classes: Assume log likelihood ratio is linear

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

\[
\text{logit}(P(C_1 \mid x)) = \log \frac{P(C_1 \mid x)}{1 - P(C_1 \mid x)} = \log \frac{p(x \mid C_1)}{p(x \mid C_2)} + \log \frac{P(C_1)}{P(C_2)}
\]

\[
= w^T x + w_0
\]

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

\[
y = \hat{P}(C_1 \mid x) = \frac{1}{1 + \exp[-(w^T x + w_0)]}
\]
Training: Two Classes

\[ \mathcal{X} = \{ x^t, r^t \} \quad r^t \mid x^t \sim \text{Bernoulli}(y^t) \quad r = 1 \text{ if } x \text{ is in } C_1 \]

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

By Bayes’ rule and algebra, we get sigmoid function

\[ I(w, w_0 \mid \mathcal{X}) = \prod_t \left( y^t \right)^{r^t} \left( 1 - y^t \right)^{1-r^t} \]

Turn Bernoulli likelihood into error function (cross-entropy)

\[ E = -\log I \]

\[ E(w, w_0 \mid \mathcal{X}) = -\sum_t r^t \log y^t + (1 - r^t) \log (1 - y^t) \]
Training: Gradient-Descent

\[ E(w, w_0 \mid X) = -\sum_t r_t \log y_t + (1 - r_t) \log (1 - y_t) \]

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

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

\[ = \eta \sum_t (r_t - y_t) x_j^t, j = 1, \ldots, d \]

\[ \Delta w_0 = -\eta \frac{\partial E}{\partial w_0} = \eta \sum_t (r_t - y_t) \]
For $j = 0, \ldots, d$
\[ w_j \leftarrow \text{rand}(-0.01, 0.01) \]
Repeat
\[ \Delta w_j \leftarrow 0 \]
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
$K > 2$ Classes

$$\mathcal{X} = \{ x^t, r^t \} \quad r^t \mid x^t \sim \text{Mult}_K (1, y^t)$$

$$\log \frac{p(x \mid C_i)}{p(x \mid C_K)} = w^T_i x + w_{i0}$$

$$y = \hat{P}(C_i \mid x) = \frac{\exp[\mathbf{w}_i^T \mathbf{x} + w_{i0}]}{\sum_{j=1}^{K} \exp[\mathbf{w}_j^T \mathbf{x} + w_{j0}]} \quad i = 1, \ldots, K$$

$$l(\{w_i, w_{i0}\}_i \mid \mathcal{X}) = \prod_t \prod_i (y_i^t)^{r_i^t}$$

$$E(\{w_i, w_{i0}\} \mid \mathbf{X}) = - \sum_t \sum_i r_i^t \log y_i^t$$

$$\Delta w_j = \eta \sum_t (r_j^t - y_j^t) x^t \quad \Delta w_{j0} = \eta \sum_t (r_j^t - y_j^t)$$
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
Generalizing the Linear Model

- Quadratic:
  \[ \log \frac{p(x | C_i)}{p(x | C_K)} = x^T W_i x + w_i^T x + w_{i0} \]

- Sum of basis functions:
  \[ \log \frac{p(x | C_i)}{p(x | C_K)} = w_i^T \phi(x) + w_{i0} \]

where \( \phi(x) \) are basis functions
  - Hidden units in neural networks (Chapters 11 and 12)
  - Kernels in SVM (Chapter 13)
Discrimination by Regression

- Classes are NOT mutually exclusive and exhaustive

\[ r^t = y^t + \varepsilon \text{ where } \varepsilon \sim \mathcal{N}(0,\sigma^2) \]

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

\[ l(w, w_0 \mid X) = \prod_t \frac{1}{\sqrt{2\pi\sigma}} \exp \left[ -\frac{(r^t - y^t)^2}{2\sigma^2} \right] \]

\[ E(w, w_0 \mid X) = \frac{1}{2} \sum_t (r^t - y^t)^2 \]

\[ \Delta w = \eta \sum_t (r^t - y^t) y^t (1 - y^t) x^t \]
Conclusion

- Linear discriminants are simple and powerful
- The basis for neural networks and vector machines
- Non-linear problems can be made linear using non-linear basis functions