EECS 545 → Machine Learning

Lecture 5 → Classification 2

- For multiclass classification, we use softmax regression
  - a generalization of logistic regression
- Logistic regression models class conditional probability as,

$$\left.\begin{aligned} p(y=1|x;w) &= \frac{\exp(w^T\phi(x))}{1+\exp(w^T\phi(x))} \\ p(y=0|x;w) &= \frac{1}{1+\exp(w^T\phi(x))} \end{aligned}\right\} \text{sum to } 1$$

- For multiclass classification, with K classes, this is modified to,

$$\left.\begin{aligned} p(y=k|x;w) &= \frac{\exp(w_k^T\phi(x))}{1+\sum_{j=1}^{K-1}\exp(w_j^T\phi(x))} \\ p(y=K|x;w) &= \frac{1}{1+\sum_{j=1}^{K-1}\exp(w_j^T\phi(x))} \end{aligned}\right\} \text{sum to } 1$$

Log-likelihood

- Setting $w_K=0$, we have,

$$p(y=k|x;w) = \frac{\exp(w_k^T\phi(x))}{\sum_{j=1}^{K}\exp(w_j^T\phi(x))} \quad \text{or} \quad p(y|x;w) = \prod_{k=1}^{K}\left[\frac{\exp(w_k^T\phi(x))}{\sum_{j=1}^{K}\exp(w_j^T\phi(x))}\right]^{y=k}$$

$$\log p(D|w) = \sum_i \log p(y^{(i)}|x^{(i)},w)$$

$$= \sum_i \log \prod_{k=1}^{K} \frac{\exp(w_k^T\phi(x))}{\sum_{k=1}^{K}\exp(w_k^T\phi(x))}$$

- Learn w by gradient ascent or Newton's method

Probabilistic Generative Models

- Goal → learn the dist^ns $p(C_k|x)$
- Discriminative models → directly model $p(C_k|x)$ and learn parameters from training set
  - Logistic Regression
  - Softmax Regression
- Generative models → learn joint densities $p(x,C_k)$ by learning $p(x|C_k)$ and priors $p(C_k)$ and then use Bayes rule for predicting class C_k given x
  - Gaussian Discriminant Analysis
  - Naive Bayes

- Bayes theorem allows us to calculate class probabilities as,

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)} = \frac{p(x|C_k)p(C_k)}{\sum_{k'} p(x|C_{k'})p(C_{k'})}$$

- To calculate the class probabilities, we need to find the probability distⁿs of $p(C_k)$ and $p(x|C_k)$

- In a two class example, we have $p(C_1|x) = \dfrac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1)+p(x|C_2)p(C_2)}$

- Use log odds $\rightarrow a = \ln\left(\dfrac{p(C_1|x)}{p(C_2|x)}\right) = \ln\left(\dfrac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)}\right)$

- Then define the posterior via the sigmoid $\rightarrow p(C_1|x) = \dfrac{1}{1+\exp(-a)} = \sigma(a)$

Gaussian Discriminant Analysis

- $p(x|C_k)$ is Gaussian distⁿ

$$p(x|C_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)\right\}$$

- $p(C_k) \rightarrow$ Bernoulli (constant)
- Basic GDA assumes same covariance for all classes
- With this assumption, $p(C_1|x) = \sigma(w^T x + w_0)$ where $w = \Sigma^{-1}(\mu_1 - \mu_2)$

$$w_0 = -\frac{1}{2}\mu_1^T \Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2^T \Sigma^{-1}\mu_2 + \log\left\{\frac{p(C_1)}{p(C_2)}\right\}$$

Derivation

$$p(x, C_1) = p(x|C_1)p(C_1)$$

$$= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)\right\} p(C_1)$$

$$P(x, C_2) = p(x|C_2)p(C_2)$$

$$= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2)\right\} p(C_2)$$

$$\log\left[\frac{p(C_1|x)}{p(C_2|x)}\right] = \log\left[\frac{p(C_1|x)}{1-p(C_1|x)}\right]$$

$$= \log\left[\frac{\exp\{-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)\}}{\exp\{-\frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2)\}}\right] + \log\left[\frac{p(C_1)}{P(C_2)}\right]$$

$$= \left\{-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)\right\} - \left\{-\frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2)\right\} + \log\left[\frac{p(C_1)}{p(C_2)}\right]$$

$$= (\mu_1 - \mu_2)^T \Sigma^{-1} x - \frac{1}{2}\mu_1^T \Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2^T \Sigma^{-1}\mu_2 + \log\left[\frac{p(C_1)}{p(C_2)}\right]$$

$$= (\Sigma^{-1}(\mu_1 - \mu_2))^T x + \omega_0 \qquad \text{where } \omega_0 = -\frac{1}{2}\mu_1^T \Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2^T \Sigma^{-1}\mu_2 + \log\left[\frac{p(C_1)}{p(C_2)}\right]$$

- Sigmoid function $\sigma(a) = \frac{1}{1+\exp(-a)}$ with previous log odds derivation generalizes to the softmax function.

$$p_i = \frac{\exp(q_i)}{\sum_j \exp(q_j)}$$

- At the decision boundary, $p(C_1|x) = p(C_2|x)$
- With the same covariance matrices the decision boundary is linear
- Decision boundary can be quadratic when each class has a different covariance

- Logistic Regression → for M dimensional feature space, the model fits M parameters
- GDA → $2M$ parameters for means for $p(x|C_1)$ and $p(x|C_2)$
  $M(M+1)/2$ parameters for the shared covariance matrix
- Logistic regression has less parameters and is more flexible about data dist^n
- GDA has a stronger modeling assumption and works well when the dist^n follows the assumption

<u>Naive Bayes</u>

- $p(C_k)$ is Bernoulli (constant)
- $p(x|C_k)$ is factorized, each coordinate of x is conditionally independent of other coordinates given the class label

$$p(x_1, \dots, x_M | C_k) = p(x_1|C_k) \dots p(x_M|C_k) = \prod_{j=1}^{M} p(x_j|C_k)$$

- For classification, use Bayes rule, $p(C_1|x) = \frac{p(C_1, x)}{p(x)} = \frac{p(C_1, x)}{p(C_1, x) + p(C_2, x)}$

- Find the class C_k that maximizes $p(C_k|x)$ using the Bayes rule,

$$\arg\max_k p(C_k|x) = \arg\max_k p(C_k|x)$$

$$= \arg\max_k p(C_k)\, p(x|C_k)$$

$$= \arg\max_k p(C_k) \prod_{j=1}^{M} p(x_j|C_k)$$

- $\prod_{j=1}^{M} p(x_j|C_k) = \prod_{j=1}^{M} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{-\frac{(x_j - \mu_j)^2}{2\sigma_j^2}\right\}$

- Generative approach is model based
  - can generate synthetic data from $p(x|C_1)$
  - we can see how good the model is when we compare the real and synthetic data
- Discriminative model typically has fewer parameters
  - less assumptions about data dist^n
  - Constructing features may require prior knowledge