Logistics

• HW 1 is due today, Sept. 23
• Midterm
  – Monday, Oct. 28
• Final exam
  – In the last class, Wed. Dec. 4
• Project report due during the final exam week
• Project milestone
  – Form teams of 2-3 people
  – Submit project proposal on Oct. 16
  – Project pitch on Oct. 21 in class
Outline

• Linear classifier
  – LDA on one dimension
  – Linear discriminant functions
  – Multi-variate LDA

• Bias-variance tradeoff
  – Derivation for linear regression
LDA

• Classify to one of k classes (multi-class)

• LDA uses Bayes Theorem to estimate it

\[- \text{P}[Y = k | X = x] = \frac{\text{P}[X = x | Y = k] \text{P}[Y = k]}{\text{P}[X = x]} \]

– Let \( \pi_k = \text{P}[Y = k] \) be the prior probability of class k and \( f_k(x) = \text{P}[X = x | Y = k] \)

• Generative model

  – Given X and Y, learns the joint probability \( P(X, Y) \)

• Discriminative model

  – Given X and Y, learns a decision function for classification
LDA

\[ \Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}. \]

Assume \( f_k(x) \) is Gaussian!

Unidimensional case (d=1)

\[ f_k(x) = \frac{1}{\sqrt{2\pi\sigma_k}} \exp \left( -\frac{1}{2\sigma_k^2} (x - \mu_k)^2 \right) \]

\[ p_k(x) = \frac{\pi_k}{\sum_{l=1}^{K} \frac{1}{\sqrt{2\pi\sigma}} \exp \left( -\frac{1}{2\sigma^2} (x - \mu_k)^2 \right)} \]

Assumption: \( \sigma_1 = \ldots \sigma_k = \sigma \)
LDA decision boundary

Pick class $k$ to maximize

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

Example: $k = 2, \pi_1 = \pi_2$

Classify as class 1 if $x > \frac{\mu_1 + \mu_2}{2}$

True decision boundary | Estimated decision boundary
LDA in practice

Given training data \((x_i, y_i), i = 1, \ldots, n, y_i \in \{1, \ldots, K\}\)

1. Estimate mean and variance

\[
\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i
\]

\[
\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2
\]

2. Estimate prior

\[
\hat{\pi}_k = \frac{n_k}{n}.
\]

Given testing point \(x\), predict \(k\) that maximizes:

\[
\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)
\]
Multi-Variate Normal

Many sample points from a multivariate normal distribution with
\[ \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \] and
\[ \Sigma = \begin{bmatrix} 1 & 3/5 \\ 3/5 & 2 \end{bmatrix}, \]
shown along with the 3-sigma ellipse, the two marginal distributions, and the two 1-d histograms.

\[ \mathbf{X} \sim \mathcal{N}(\mu, \Sigma), \]

with \( k \)-dimensional mean vector
\[ \mu = \mathbb{E}[\mathbf{X}] = [\mathbb{E}[X_1], \mathbb{E}[X_2], \ldots, \mathbb{E}[X_k]]^T, \]

and \( k \times k \) covariance matrix
\[ \Sigma_{i,j} =: \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] = \text{Cov}[X_i, X_j] \]
\[ \Sigma =: \mathbb{E}[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = [\text{Cov}[X_i, X_j]; 1 \leq i, j \leq k]. \]

\[ f_k(x) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} e^{-\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)}. \]
Multi-variate LDA

\[ \Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}. \]

Assume \( \Sigma_k = \Sigma \)

\[
\log \frac{\Pr(Y = k|X = x)}{\Pr(Y = l|X = x)} = \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l} \\
= \log \frac{\pi_k}{\pi_l} - \frac{1}{2} (\mu_k + \mu_l)^T \Sigma^{-1} (\mu_k - \mu_l) \\
+ x^T \Sigma^{-1} (\mu_k - \mu_l),
\]

Linear decision boundary between classes \( k \) and \( l \)

Linear discriminant functions

\[ \delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k \]

Given \( x \), classify to class \( k \): \( \text{argmax}_k \delta_k(x) \)
Example 3 classes

3 Normal distributions with same co-variance, but different means

LDA decision boundary
Multi-variate LDA

Given training data \((x_i, y_i), i = 1, ..., n, y_i \in \{1, ..., K\}\)

1. Estimate mean and variance

- \(\hat{\pi}_k = N_k/N\), where \(N_k\) is the number of class-\(k\) observations;
- \(\hat{\mu}_k = \sum_{i : y_i = k} x_i / N_k\);
- \(\hat{\Sigma} = \sum_{k=1}^{K} \sum_{i : y_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / (N - K)\).

2. Estimate prior

Given testing point \(x\), predict \(k\) that maximizes:

\[
\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k
\]
Generalization in ML

- Goal is to generalize well on new testing data
- Risk of overfitting to training data
  - MSE close to 0, but performs poorly on test data
Bias = Difference between estimated and true models
Variance = Model difference on different training sets

MSE is proportional to Bias + Variance
Bias-Variance Decomposition

- Assume that \( y = f(x) + \epsilon \)
  - Noise \( \epsilon \) is sampled from a normal distribution with 0 mean and variance \( \sigma^2 \): \( \epsilon \sim N(0, \sigma^2) \)
  - Noise lower-bounds the performance we can achieve

- Recall the following objective function:
  \[
  J(\theta) = \sum_{i=1}^{n} (h_\theta(x_i) - y_i)^2
  \]

- We can re-write this as the expected value of the squared error: \( \mathbb{E} \left( y - h_\theta(x) \right)^2 \)

\( f(x) \): True model (not necessarily linear)
\( h_\theta(x) \): Learned model (linear)
Bias-Variance Decomposition

\[
E[(y - h_\theta(x))^2] = E[(y - f(x) + f(x) - h_\theta(x))^2] \\
= E[(y - f(x))^2] + E[(f(x) - h_\theta(x))^2] \\
+ 2E[(f(x) - h_\theta(x))(y - f(x))] \\
= E[(y - f(x))^2] + E[(f(x) - h_\theta(x))^2] \\
+ 2(E[f(x)h_\theta(x)] + E[yf(x)] - E[yh_\theta(x)] - E[f(x)^2])
\]

cancels
cancels

Therefore,

\[
E[(y - h_\theta(x))^2] = E[(y - f(x))^2] + E[(f(x) - h_\theta(x))^2] \\
= E[\epsilon^2] + E[(f(x) - h_\theta(x))^2]
\]

Aside:

Definition of Variance

\[
\text{var}(z) = E[(z - E[z])^2]
\]

This is actually \(\text{var}(\epsilon)\), since mean is 0
Bias-Variance Decomposition

\[ E[(y - h_\theta(x))^2] = \text{var}(\epsilon) + E[(f(x) - h_\theta(x))^2] \]
\[
= \text{var}(\epsilon) + E[(f(x) - E[h_\theta(x)])^2] + E[E[h_\theta(x)] - h_\theta(x)]^2 \\
+ 2E[(E[h_\theta(x)] - h_\theta(x))(f(x) - E[h_\theta(x)])] \\
= \text{var}(\epsilon) + E[(f(x) - E[h_\theta(x)])^2] + E[(E[h_\theta(x)] - h_\theta(x))^2] \\
+ 2(E[f(x)E[h_\theta(x)]] - E[E[h_\theta(x)]^2] - E[f(x)h_\theta(x)] + E[h_\theta(x)E[h_\theta(x)]]
\]

cancels cancels

Therefore,

\[ E[(y - h_\theta(x))^2] = \text{var}(\epsilon) + E[(f(x) - E[h_\theta(x)])^2] + E[(E[h_\theta(x)] - h_\theta(x))^2] \]

\[ E[(y - h_\theta(x))^2] = \text{bias}(h_\theta(x))^2 + \text{var}(h_\theta(x)) + \sigma^2 \]
Review

• LDA
  – Example of generative model
  – Use Bayes Theorem to estimate the probability that label is from each class k
  – Assumes normal distribution of features in each class

• Bias-Variance tradeoff for linear regression
  – Decompose expectation of testing error as a sum of squared bias, variance, and noise term
  – Shows that bias and variance need to be simultaneously minimized
Acknowledgements

• Slides made using resources from:
  – Andrew Ng
  – Eric Eaton
  – David Sontag

• Thanks!