Announcements

- Staff mailing list: cs195-5tas@cs.brown.edu
- Problem set automatic submission: see website for corrected instructions.
Review

- Geometry of multivariate Gaussians
- Classification via direct regression
- Geometry of projections; linear decision boundaries.
Geometry of projections

- \( \mathbf{w}^T \mathbf{x} = 0 \): a line passing through the origin and orthogonal to \( \mathbf{w} \)

- \( \mathbf{w}^T \mathbf{x} + w_0 = 0 \) shifts the line along \( \mathbf{w} \).
Geometry of projections

\[ w^T x = 0 \]: a line passing through the origin and orthogonal to \( w \)

\[ w^T x + w_0 = 0 \] shifts the line along \( w \).
Geometry of projections

- $\mathbf{w}^T \mathbf{x} = 0$: a line passing through the origin and **orthogonal** to $\mathbf{w}$
- $\mathbf{w}^T \mathbf{x} + w_0 = 0$ shifts the line along $\mathbf{w}$.
Geometry of projections

- $w^T x = 0$: a line passing through the origin and orthogonal to $w$
- $w^T x + w_0 = 0$ shifts the line along $w$. 

$$w_0 + w^T x = 0$$
Geometry of projections

- $w^T x = 0$: a line passing through the origin and orthogonal to $w$
- $w^T x + w_0 = 0$ shifts the line along $w$.

- $x'$ is the projection of $x$ on $w$. 

$-\frac{w_0}{\left\| w \right\|}$

$\frac{w_0 + w^T x_0}{\left\| w \right\|}$

$w_0 + w^T x = 0$
Geometry of projections

- $\mathbf{w}^T \mathbf{x} = 0$: a line passing through the origin and orthogonal to $\mathbf{w}$
- $\mathbf{w}^T \mathbf{x} + w_0 = 0$ shifts the line along $\mathbf{w}$.

- $\mathbf{x}'$ is the projection of $\mathbf{x}$ on $\mathbf{w}$.

- Set up a new 1D coordinate system: $\mathbf{x} \rightarrow (w_0 + \mathbf{x}^T \mathbf{x})/\|\mathbf{w}\|$.
Consider a scalar projection

$$f : \mathbf{x} \rightarrow w_0 + \mathbf{w}^T \mathbf{x}$$

We can study how well the projected values corresponding to different classes are separated

- This is a function of $\mathbf{w}$; some projections may be better than others.
Distribution in 1D projection

• Consider a scalar projection

\[ f : \mathbf{x} \rightarrow w_0 + \mathbf{w}^T \mathbf{x} \]

• We can study how well the projected values corresponding to different classes are separated

  – This is a function of \( \mathbf{w} \); some projections may be better than others.
Distribution in 1D projection

- Consider a scalar projection

\[ f : \mathbf{x} \rightarrow w_0 + \mathbf{w}^T \mathbf{x} \]

- We can study how well the projected values corresponding to different classes are separated
  - This is a function of \( \mathbf{w} \); some projections may be better than others.
Linear discriminant and dimensionality reduction

The discriminant function $f(x; w) = w_0 + w^T x$ reduces the dimension of examples from $d$ to 1; the components orthogonal to $w$ become irrelevant.

$$\hat{y} = +1 \iff f(x; w) > 0$$
Projections and classification

What objective are we optimizing the 1D projection for?
Objective: class separation

- We want to minimize “overlap” between projections of the two classes.
- An obvious idea: maximize separation between the projected means
Separation of the means

- \( N_{+1} \) examples of class +1, \( N_{-1} \) examples of class -1.

- The empirical mean of each class:

\[
\begin{align*}
\mathbf{m}_{+1} &= \frac{1}{N_{+1}} \sum_{y_i=+1} x_i, \\
\mathbf{m}_{-1} &= \frac{1}{N_{-1}} \sum_{y_i=-1} x_i
\end{align*}
\]

- We can look for projection \( \hat{\mathbf{w}} \) such that

\[
\hat{\mathbf{w}} = \arg \max_{\mathbf{w}} \mathbf{w}^T (\mathbf{m}_{+1} - \mathbf{m}_{-1})
\]
The effect of scaling $w$

- What happens if we multiply $w_0, w$ by some $a > 1$?
The effect of scaling $w$

- What happens if we multiply $w_0, w$ by some $a > 1$?

  $$\text{sign}(aw_0 + aw^T x) = \text{sign} \left( a(w_0 + w^T x) \right) = \text{sign}(w_0 + w^T x),$$

  so the decision boundary doesn’t change.

- However, $aw^T(m_{+1} - m_{-1}) > w^T(m_{+1} - m_{-1})$!

- So, $w^T(m_{+1} - m_{-1})$ is unbounded. We need to impose an additional constraint: $\|w\| = 1$.

- The solution, not surprisingly, is

  $$\arg\max_{\|w\|=1} w^T(m_{+1} - m_{-1}) = \frac{m_{+1} - m_{-1}}{\|m_{+1} - m_{-1}\|}$$
Separation of the means: example

\[ \hat{w} = \underset{\|w\|=1}{\text{argmax}} w^T (m_+ - m_-) \]

- Also want to make projection of each class “compact”...
1D projections of Gaussians

As an illustration, consider 1D projections of a multivariate Gaussian.

- Let $p(x) = \mathcal{N}(x; \mu, \Sigma)$.

- For any $A$, $p(Ax) = \mathcal{N}(Ax; A\mu, A\Sigma A^T)$.

- To get the marginal of 1D projection on the direction defined by a unit vector $v$:
  - Make $R$ a rotation such that $R[1, 0, \ldots, 0]^T = v$, and take the first coordinate of $Rx$.
  - Alternatively, directly compute $\Sigma_v = v^T \Sigma v$; that’s the variance of the marginal.

- Matlab demo: margGausDemo.m
Fisher’s linear discriminant analysis

- Criterion to be maximized:

\[ J_{Fisher}(w) = \frac{\text{separation between projected means}^2}{\text{sum of projected within-class variances}} \]

- Numerator: *between-class scatter* \( (w^T(m_{+1} - m_{-1}))^2 \)

- Denominator: *within-class scatter* \( w^T(N_{-1}S_{-1} + N_{+1}S_{+1})w \), where

\[ S_c = \frac{1}{N_c} \sum_{y_i = c} (x_i - m_c)(x_i - m_c)^T. \]

  - The denominator is the sum of estimated 1D class covariances, after data are projected to \( w \), weighted by number of samples in each class.
Fisher’s LDA

\[
J_{Fisher}(\mathbf{w}) = \frac{(\mathbf{w}^T(\mathbf{m}_+ - \mathbf{m}_-))^2}{\mathbf{w}^T(\mathbf{N}_-\mathbf{S}_- + \mathbf{N}_+\mathbf{S}_+) \mathbf{w}}
\]

- **Best 1D projection:** \( \hat{\mathbf{w}} = \arg\max_{\mathbf{w}} J_{Fisher}(\mathbf{w}) \)

- **Setting the derivative of** \( J \) w.r.t. \( \mathbf{w} \) to zero, get solution:

\[
\hat{\mathbf{w}} \propto (\mathbf{N}_-\mathbf{S}_- + \mathbf{N}_+\mathbf{S}_+)^{-1}(\mathbf{m}_+ - \mathbf{m}_-)
\]

Notation: \( \propto \) means “proportional to”, up to a constant factor.
Classification in 1D

\[ \hat{y} = \text{sign} \left( w_0 + w^T x \right) \]

- Once we have \( w \) how do we set \( w_0 \)?

- Possibility 1, with no explicit probabilistic assumptions:
  - find threshold that minimizes empirical classification error.

- Possibility 2: Make assumptions about data distribution, and derive a theoretically optimal decision boundary.
  - This has been studied in decision theory.
Decision theory

- Suppose we are working under the 0/1 loss:

\[
L(\hat{y}, y) = \begin{cases} 
1 & \text{if } \hat{y} \neq y, \\
0 & \text{if } \hat{y} = y.
\end{cases}
\]

- Then the expected loss of a \(C\)-way classifier \(h(x)\) is

\[
E_{p(y|x)p(x)} [L(h(x), y)] = \int_x \left[ \sum_{c=1}^{C} \Pr(y = c, h(x) \neq c \mid x) \right] p(x) dx.
\]

- Enough to minimize \(\Pr(h(x) \neq y|x)\) for each \(x\).

- In other words, need to minimize probability of classification error.
Minimizing expected classification error

- Suppose $C = 2$. We want to minimize $\Pr(h(x) \neq y|x)$. There are two disjoint “bad” events:
  - $E_1$: real label is $+1$, we classify as $-1$
  - $E_2$: real label is $-1$, we classify as $+1$
Suppose \( C = 2 \). We want to minimize \( \Pr(h(x) \neq y|x) \). There are two disjoint “bad” events:

- \( E_1 \): real label is +1, we classify as −1
- \( E_2 \): real label is −1, we classify as +1

\[
\Pr(h(x) \neq y|x) = \Pr(y = +1, h(x) = -1 | x) + \Pr(y = -1, h(x) = +1 | x).
\]
Minimizing expected classification error

• Suppose $C = 2$. We want to minimize $\Pr(h(x) \neq y|x)$. There are two disjoint “bad” events:
  
  – $E_1$: real label is $+1$, we classify as $-1$
  – $E_2$: real label is $-1$, we classify as $+1$

\[
\Pr(h(x) \neq y|x) = \Pr(y = +1, h(x) = -1|x) + \Pr(y = -1, h(x) = +1|x).
\]

⇒ Decision rule that minimizes probability of both $E_1$ and $E_2$:

\[
h^*(x) = \arg\max_c p(c|x).
\]

• Holds for $C > 2$ as well!
Bayes rule

- Some terminology:

  class-conditional density \( p_c(x) = p(x | y = c) \)

  prior probability \( P_c = p(y = c) \)

  posterior probability \( p(y = c | x) \)

  data probability \( p(x) = \sum_c p(x, y = c) = \sum_c p(x | y = c) P_c \).

- Usually we don’t have direct access to \( \Pr(y | x) \). But suppose we know \( p(x | y), p(y) \).

- Bayes rule: Using the product rule \( p(a, b) = p(a | b)p(b) = p(b | a)p(a) \),

  \[
p(y | x) = \frac{p(x | y)p(y)}{p(x)}.
  \]
Next time

Finish talking about decision theory.
Generative models for classification.
Density estimation.