Administrivia

• HW4 out
  ‣ based on feedback survey,
    ‣ fewer questions: 4, but only do 3
    ‣ range of problem types: focus on those that help your understanding
    ‣ split out “spoilers” for Q2

• Midterm
  ‣ mean 65 (out of 95), std dev 11.3
  ‣ back at end of class
Review

- Cone & QP duality
  - \( \min c^T x + x^T H x / 2 \)  s.t.  \( A x + b \in K \quad x \in L \)
  - \( \max -z^T H z / 2 - b^T y \)  s.t.  \( H z + c - A^T y \in L^* \quad y \in K^* \)

- KKT conditions
  - primal:  \( A x + b \in K \quad x \in L \)
  - dual:  \( H z + c - A^T y \in L^* \quad y \in K^* \)
  - quadratic:  \( H x = H z \)
  - comp. slack:  \( y^T (A x + b) = 0 \quad x^T (H z + c - A^T y) = 0 \)
Support vector machines

(review)

Figure 4, Two dimensional representation from MVU of space, as discussed in Fig. 2.

The circular arrangement reflects the underlying rotational degree of freedom in the data set. Note posing representative images on top of their corresponding presentation discovered by MVU is easily visualized by superimposing images of the same person's face in different poses and expressions.

Each image was grayscale and contained pixel information so that the pixel space had dimensionality $d = 1960$. Figure 5 shows a three dimensional representation of face images of a teapot viewed from different angles in the plane. The teapot was viewed under a full 491 degrees of rotation. Each image of the teapot viewed from different angles in the plane.

In this representation, the number of dimensions needed to preserve local distances while maximizing variance is reflected in the cyclic degree of freedom in the data set. Note how this representation supports judgments of similarity reflecting the cyclic degree of freedom in the data set.

The two dimensional representation is easier to visualize. The data set consisted of 3116 pixels that are particularly easy to visualize. We have used maximum variance unfolding (MVU) to analyze many high dimensional data sets of interest. Here we show some solutions.

In particular, it can be solved efficiently with polynomial time guarantees and many off-the-shelf solvers are available in the public domain. Support vector machines (SVM) are convex. In particular, it can be solved efficiently with many off-the-shelf solvers are available in the public domain.

Fig. 5 shows a three dimensional representation of face images of a teapot viewed from different angles in the plane.

Figure 6 shows a two dimensional representation of words discovered by MVU.

Weinberger et al. developed methods for visualizing high dimensional data, such as co-occurrence statistics of the most frequently occurring words in a large corpus of text. Each word was represented by a sparse dimensional vector of normalized counts, as typically collected for bigram language modeling. The figure shows that many semantic relationships between words are preserved despite the drastic reduction in dimensionality from linear versus nonlinear dimensionality reduction.

The inputs to MVU were derived from co-occurrence statistics of the most frequently occurring words in a large corpus of text. Each word was represented by a sparse dimensional vector of normalized counts, as typically collected for bigram language modeling.

The figure shows that many semantic relationships between words are preserved despite the drastic reduction in dimensionality from linear versus nonlinear dimensionality reduction.

In this representation, the number of dimensions needed to preserve local distances while maximizing variance is reflected in the cyclic degree of freedom in the data set.

The circular arrangement reflects the underlying rotational degree of freedom in the data set. Note posing representative images on top of their corresponding presentation discovered by MVU is easily visualized by superimposing images of the same person's face in different poses and expressions.

Each image was grayscale and contained pixel information so that the pixel space had dimensionality $d = 1960$. Figure 5 shows a three dimensional representation of face images of a teapot viewed from different angles in the plane.

In this representation, the number of dimensions needed to preserve local distances while maximizing variance is reflected in the cyclic degree of freedom in the data set. Note how this representation supports judgments of similarity reflecting the cyclic degree of freedom in the data set.

The two dimensional representation is easier to visualize. The data set consisted of 3116 pixels that are particularly easy to visualize. We have used maximum variance unfolding (MVU) to analyze many high dimensional data sets of interest. Here we show some solutions.

In particular, it can be solved efficiently with many off-the-shelf solvers are available in the public domain. Support vector machines (SVM) are convex. In particular, it can be solved efficiently with many off-the-shelf solvers are available in the public domain.

Fig. 5 shows a three dimensional representation of face images of a teapot viewed from different angles in the plane.

Figure 6 shows a two dimensional representation of words discovered by MVU.

Weinberger et al. developed methods for visualizing high dimensional data, such as co-occurrence statistics of the most frequently occurring words in a large corpus of text. Each word was represented by a sparse dimensional vector of normalized counts, as typically collected for bigram language modeling. The figure shows that many semantic relationships between words are preserved despite the drastic reduction in dimensionality from linear versus nonlinear dimensionality reduction.

The inputs to MVU were derived from co-occurrence statistics of the most frequently occurring words in a large corpus of text. Each word was represented by a sparse dimensional vector of normalized counts, as typically collected for bigram language modeling.

The figure shows that many semantic relationships between words are preserved despite the drastic reduction in dimensionality from linear versus nonlinear dimensionality reduction.

In this representation, the number of dimensions needed to preserve local distances while maximizing variance is reflected in the cyclic degree of freedom in the data set. Note posing representative images on top of their corresponding presentation discovered by MVU is easily visualized by superimposing images of the same person's face in different poses and expressions.

Each image was grayscale and contained pixel information so that the pixel space had dimensionality $d = 1960$. Figure 5 shows a three dimensional representation of face images of a teapot viewed from different angles in the plane.

In this representation, the number of dimensions needed to preserve local distances while maximizing variance is reflected in the cyclic degree of freedom in the data set. Note how this representation supports judgments of similarity reflecting the cyclic degree of freedom in the data set.

The two dimensional representation is easier to visualize. The data set consisted of 3116 pixels that are particularly easy to visualize. We have used maximum variance unfolding (MVU) to analyze many high dimensional data sets of interest. Here we show some solutions.

In particular, it can be solved efficiently with many off-the-shelf solvers are available in the public domain. Support vector machines (SVM) are convex. In particular, it can be solved efficiently with many off-the-shelf solvers are available in the public domain.

Fig. 5 shows a three dimensional representation of face images of a teapot viewed from different angles in the plane.

Figure 6 shows a two dimensional representation of words discovered by MVU.

Weinberger et al. developed methods for visualizing high dimensional data, such as co-occurrence statistics of the most frequently occurring words in a large corpus of text. Each word was represented by a sparse dimensional vector of normalized counts, as typically collected for bigram language modeling. The figure shows that many semantic relationships between words are preserved despite the drastic reduction in dimensionality from linear versus nonlinear dimensionality reduction.

The inputs to MVU were derived from co-occurrence statistics of the most frequently occurring words in a large corpus of text. Each word was represented by a sparse dimensional vector of normalized counts, as typically collected for bigram language modeling.

The figure shows that many semantic relationships between words are preserved despite the drastic reduction in dimensionality from linear versus nonlinear dimensionality reduction.
Support vector machines

10-725 Optimization
Geoff Gordon
Ryan Tibshirani
### SVM duality

- \[ \min ||v||^2/2 + \sum s_i \quad \text{s.t.} \quad y_i (x_i^Tv - d) \geq 1 - s_i \quad s_i \geq 0 \]

- \[ \min v^Tv/2 + l^Ts \quad \text{s.t.} \quad Av - yd + s - l \geq 0 \]

\[ A = \begin{pmatrix} y_i & x_i^T \end{pmatrix} \]

\[ \alpha^T(Av - yd + s - l) \geq 0 \]

\[ \text{obj} \geq v^Tv/2 + l^Ts - \alpha^T(Av - yd + s - l) - tr^s \]

\[ \begin{align*}
\min \left[ \begin{array}{c}
\alpha^T A^T \alpha / 2 + 1^T \alpha \\
v^T v / 2 + l^T s - \alpha^T (Av - yd + s - l) - tr^s
\end{array} \right]
\end{align*} \]

\[ \nabla_v : \quad 0 = v - A^T \alpha \Rightarrow v = A^T \alpha \]

\[ \nabla_d : \quad 0 = \alpha^T y \]

\[ \nabla_s : \quad 0 = 1 - \alpha - t \quad 1 = \alpha + t \]

\[ \nabla_l : \quad 0 = y^T d \]

\[ \alpha \geq 0 \]

\[ \alpha \leq 1 \]
Interpreting the dual

\[ \text{max } 1^T \alpha - \alpha^T K \alpha / 2 \quad \text{s.t.} \quad y^T \alpha = 0 \quad 0 \leq \alpha \leq 1 \]

\( \alpha: \) example weights

\( \alpha > 0: \) support vectors

\( \alpha < 1: \) slack = 0

\( y^T \alpha = 0: \) \( \sum y_i \alpha_i = 0 \)

\( \sum \alpha_i = \alpha_{\text{tot}} \)
From dual to primal

- \( \max 1^T \alpha - \alpha^T K \alpha / 2 \) s.t. \( y^T \alpha = 0 \) \( 0 \leq \alpha \leq 1 \)

\[
V = \sum_{\alpha_i \neq 0} \alpha_i \\
= \sum_i y_i \cdot x_i \alpha_i \\
= \sum_i \frac{x_i \alpha_i}{\sum_{\alpha_j \neq 0} \alpha_j} - \sum_i \frac{x_i \alpha_i}{\sum_{\alpha_j \neq 0} \alpha_j}
\]

\( x_i > 0 \) \\
\( y_i \cdot x_i \cdot \alpha_i - y_i \cdot \alpha_i = 1 - s_i \)

Geoff Gordon—10-725 Optimization—Fall 2012
A suboptimal support set
Why is the dual useful?

\[
\max l^T \alpha - \alpha^T K \alpha / 2 \quad \text{s.t.} \quad y^T \alpha = 0 \quad 0 \leq \alpha \leq 1
\]

• SVM: n examples, m features: \( x_i = \phi(u_i) \in \mathbb{R}^m \)
  
  ‣ primal: \( n \) vars \( n \) constrs
  
  ‣ dual: \( n \) vars 1 equality \( 2n \) box
The kernel trick

• Don’t even need to know features $x_i = \phi(u_i)$, as long as we can compute dot products $x_i^T x_j$

• Matrix of dot products:
  - $K_{ij} = x_i^T x_j = \phi(u_i)^T \phi(u_j) = k(u_i, u_j)$
  - only need subroutine for $k$ (don’t care about $\phi$)
  - how do we know $k$ works?
    - $K \succeq 0 \quad K = K^T \quad \forall \exists u_i \exists_k = 1..n$
  - this is a “positive definite function,” aka “Mercer kernel”—exists many examples
Examples of kernels

- \( K(u_i, u_j) = (1 + u_i^T u_j)^d \)
  - can represent any degree-\( d \) polynomial
  - i.e., decision surface is \( p(u) = b \) for degree-\( d \) poly \( p \)

- \( K(u_i, u_j) = (u_i^T u_j)^d \)
  - polynomial where all terms have degree exactly \( d \)
  - \( d=1 \) reduces to original (linear) SVM

- \( K(u_i, u_j) = \exp(-||u_i - u_j||^2/2\sigma^2) \)
  - Gaussian radial basis functions of width \( \sigma \)
Gaussian kernel

$\sigma = 0.5$
Interior-point methods

10-725 Optimization
Geoff Gordon
Ryan Tibshirani
**Ball center**

*aka Chebyshev center*

- $X = \{ x | Ax + b \geq 0 \}$

- **Ball center:**
  - $\max_x \min_i \text{dist} (x, \frac{a_i^* x + b_i^*}{\|a_i\|} = 0)$
  - if $\|a_i\| = 1$
    - $\max_{x_j^*} t \; s.t. \; s:Ax + b \geq t1$
  - in general:
    - $\frac{s_i}{\|a_i\|} \geq t \; \forall i$
**Ellipsoid center**

*aka max-volume inscribed ellipsoid*

- Center $d$ of largest inscribed ellipsoid
  - $E = \{ Bu + d \mid ||u||_2 \leq 1 \}$
  - $\text{vol}(E) \geq \text{vol}(X)/n$ in $\mathbb{R}^n$

- $\min \log \det B^{-1}$ s.t.
  - $a_i^T(Bu+d) + b_i \geq 0 \quad \forall i \quad \forall u \text{ with } ||u|| \leq 1$
  - $B \succeq 0$

- Convex optimization, but relatively expensive:
  - convex objective, semidefinite constraint
  - each $(u, a_i, b_i)$ yields a linear constraint on $B, d$
Analytic center

- Let $s = Ax + b$
- Analytic center:
  - $\max_{x, s \geq 0} \sum_i s_i / \|a_i\|$
  - $\min_{x, s \geq 0} - \sum \ln s_i / \|a_i\|$
  - $\geq - \sum \ln s_i + \sum \ln \|a_i\|$
Bad conditioning? No problem.

\[ a_i^T x + b_i \geq 0 \quad \min -\sum \ln(a_i^T x + b_i) \]

\[ -\sum \frac{1}{a_i^T x + b_i} a_i \]

\[ y = Mx + q \]
Newton for analytic center

- \( f(x) = -\sum \ln(a_i^T x + b_i) \)
  - \( \frac{df}{dx} = -\sum a_i / (a_i^T x + b_i) = -\sum \frac{1}{s_i} a_i = -A^T \frac{1}{s} \)
  - \( \frac{d^2f}{df^2} = \sum_i \frac{1}{(a_i^T x + b_i)^2} a_i a_i^T = A^T S^{-2} A \)

\( s = A x + b \quad S = d \cdot \gamma(s) \)
Adding an objective

• Analytic center was for: find $x$ st $Ax + b \geq 0$

• Now: $\min c^T x$ st $Ax + b \geq 0$

• Same trick:
  ‣ $\min f_t(x) = c^T x - (1/t) \sum \ln(a_i^T x + b_i)$
    ‣ parameter $t > 0$
  ‣ central path =
    ‣ $t \to 0$: analytic center
    ‣ $t \to \infty$: LP opt