Welcome back.
Projects comments available on Glookup!
Turn in homework!
I am away April 15-20.
Midterm out when I get back.
Few days takehome.
Shiftable.
Have handle on projects before that.
Progress report due Monday.

Two populations.
DNA data:
- Human 1: A ··· C ··· T ··· A
- Human 2: C ··· C ··· A ··· T
- Human 3: A ··· G ··· T ··· T

Single Nucleotide Polymorphism.
Same population?
Model: same population breeds.
Population 1: snp 843: Pr[A] = .4 , Pr[T] = .6
Population 2: snp 843: Pr[A] = .6 , Pr[T] = .4

Individual: x₁, x₂, x₃, ..., xₙ
Which population?
Comment: snps could be movie preferences, populations could be
types.
E.g., republican/democrat, shopper/saver.

Gaussians
Population 1: Gaussian with mean µ₁ ∈R², variance σ in each dim.
Population 2: Gaussian with mean µ₂ ∈R², variance σ in each dim.

Difference between humans σ per snp.
 Difference between populations ε per snp.
How many snps to collect to determine population for individual x?
x in population 1.
E[(x - µ₁)²] = dσ²
E[(x - µ₂)²] ≥ (d - 1)σ² + (µ₁ - µ₂)².
If (µ₁ - µ₂)² = dε² >> σ², then different.
→ take d >> σ²/ε²

Variance of estimator?
Roughly dε².
Signal is difference between expectations.
roughly dε²
Signal >> Noise. ↔ dε² >> √dσ².

Projection
Population 1: Gaussian with mean µ₁ ∈R², variance σ in each dim.
Population 2: Gaussian with mean µ₂ ∈R², variance σ in each dim.

Difference between humans σ per snp.
 Difference between populations ε per snp.
Project x onto unit vector v in direction µ₂ - µ₁.
E[((x - µ₁) · v)²] = 0 if x is population 1.
E[((x - µ₂) · v)²] ≥ (µ₁ - µ₂)² if x is population 2.
Std deviation is σ²! versus √dσ²!
No loss in signal!

dε² >> σ².
→ d >> σ²/ε²
Versus d >> σ²/ε⁴.
A quadratic difference in amount of data!

Which population?
Population 1: snp 843: Pr[A] = .4 , Pr[T] = .6
Population 2: snp 843: Pr[A] = .6 , Pr[T] = .4

Individual: x₁, x₂, x₃, ..., xₙ.
Population 1: snp i: Pr[A] = .6 , Pr[T] = .4
Population 2: snp i: Pr[A] = .4 , Pr[T] = .6

Simpler Calculation:
Population 1: Gaussian with mean µ₁ ∈R², variance σ in each dim.
Population 2: Gaussian with mean µ₂ ∈R², variance σ in each dim.

Don’t know much about...

Population 1: snp 843: Pr[A] = .4 , Pr[T] = .6
Population 2: snp 843: Pr[A] = .6 , Pr[T] = .4

Individual: x₁, x₂, x₃, ..., xₙ.
Population 1: snp i: Pr[xᵢ = 1] = p⁽¹⁾
Population 2: snp i: Pr[xᵢ = 0] = p⁽²⁾

Don’t know µ₁ or µ₂?

E.g., republican/democrat, shopper/saver.
Without the means?
Sample of \( n \) people.
Some (say half) from population 1,
some from population 2.

Which are which?

Near Neighbors Approach
Compute Euclidean distance squared.
Cluster using threshold.

Signal \( \mathbb{E}[d(x_1, x_2)] - \mathbb{E}[d(x_1, y_1)] \)
should be larger than noise in \( d(x, y) \)

Where \( x \)'s from one population, \( y \)'s from other.

Signal is proportional \( d^2 \).
Noise is proportional to \( \sqrt{\sigma^2} \).

\( d >> \sigma^4/\epsilon^4 \) → same type people closer to each other.

\( d >> (\sigma^4/\epsilon^4) \log n \) suffices for threshold clustering.
\( \log n \) factor for union bound over \( \binom{n}{2} \) pairs.

Best one can do?

PCA calculation.
Matrix \( A \) where rows are points.
First eigenvector of \( B = A^T A \) is maximum variance direction.

\( Av \) are projections onto \( v \).
\( vBv = (vA)^T (Av) \) is \( \Sigma (x \cdot v)^2 \).

First eigenvector, \( v \), of \( B \) maximizes \( x^T Bx \).

\( Bv = \lambda v \) for maximum \( \lambda \).

\( \rightarrow vBv = \lambda \) for unit \( v \).

Eigenvectors form orthonormal basis.
Any other vector \( av + x, x \cdot v = 0 \)

\( x \) is composed of possibly smaller eigenvalue vectors.

\( \rightarrow vBv \geq (av + x)B(av + x) \) for unit \( v \), \( av + x \).

Principal components analysis.

Remember Projection!

\[ 
\begin{array}{c}
\bullet \\
\bullet \\
\bullet
\end{array}
\]

Don't know \( \mu_1 \) or \( \mu_2 \)?

Principal component analysis:
Find direction, \( v \), of maximum variance.
Maximize \( \sum (x \cdot v)^2 \) (zero center the points)
Recall: \( (x \cdot v)^2 \) could determine population.

Typical direction variance. \( nd^2 \).

Direction along \( \mu_1 - \mu_2 \),

\[ = n(\mu_1 - \mu_2)^2 \]
\[ = nd^2 \cdot \]

Need \( d >> \sigma^2/\epsilon^2 \) at least.

When will PCA pick correct direction with good probability?
Union bound over directions. How many directions?

Infinity and beyond?

Computing eigenvalues.

Power method:
Choose random \( x \).
Repeat: Let \( x = Bx \). Scale \( x \) to unit vector.

\[ x = a_1 v_1 + a_2 v_2 + \cdots \]

\[ x_0 = B'x = a_1 \lambda_1 v_1 + a_2 \lambda_2 v_2 + \cdots \]

Mostly \( v_i \) after a while since \( \lambda_1 >> \lambda_2 \).

Cluster Algorithm:
Choose random partition.
Repeat: Compute means of partition. Project, cluster.

Choose random \( +1/-1 \) vector. Multiply by \( A^T \) (direction between means), multiply by \( A \) (project points), cluster (round to \( +1/-1 \) vector.)

Sort of repeatedly multiplying by \( AA^T \). Power method.

Nets

\( \delta \cdot \text{Net} \).
Set \( \mathcal{G} \) of directions
where all others, \( v \), are close to \( x \in \mathcal{G} \).

\( x \cdot v \geq 1 - \delta \).

\( \delta \)-Net:
\[ [\cdots, \delta/d, \cdots] \] integers \( i \) \( \in [-d/\delta, d\delta] \).

Total of \( N = (\delta/d)^{(d)} \) vectors in net.

Signal \( >> \) Noise times \( \log N = O(d \log \frac{d}{\delta}) \) to isolate direction.

\( \log N \) is due to union bound over vectors in net.

Signal (exp. projection): \( \alpha nd^2 \).
Noise (std dev.): \( \sqrt{\sigma^2} \).

\( nd >> (\sigma^4/\epsilon^4) \log d \) and \( d >> \sigma^2/\epsilon^2 \) works.

Nearest neighbor works with very high \( d >> \sigma^4/\epsilon^4 \).

PCA can reduce \( d \) to “knowing centers” case, with reasonable number of sample points.

Sum up.

Clustering mixture of gaussians.
Near Neighbor works with sufficient data.
Projection onto subspace of means is better.
Principal compent analysis can find subspace of means.
Power method computes principal component.
Generic clustering algorithm is rounded version of power method.
See you on Thursday.