Recitation 2: Linear Algebra Review

1 Notation

- $\langle u, v \rangle = u \cdot v = u^T v = \sum_{i=1}^{n} u_i v_i$
- $||u|| = \sqrt{\langle u, u \rangle}$, at least for the purposes of these notes. This is the ℓ_2 norm.

2 Warmup

- **Definition:** For the angle θ_{uv} between u and v, $cos\theta_{uv} = \frac{\langle u,v \rangle}{||u||\ ||v||}$
- **Question:** Show that $-1 \leq corr(x, y) \leq 1$
- Hint: Let $\tilde{x} = x - \bar{x}$, and $\tilde{y} = y - \bar{y}$

Remember that $corr(x, y) = \frac{cov(x,y)}{sd(x)sd(y)}$

By the definitions above, $sd(x) = \frac{1}{\sqrt{n}}||\tilde{x}||$, $sd(y) = \frac{1}{\sqrt{n}}||\tilde{y}||$, $cov(x) = \frac{1}{n}\langle \tilde{x}, \tilde{y} \rangle$

$$\begin{aligned} -1 \leq cos\theta_{\tilde{x}\tilde{y}} \leq 1 \\ -1 \leq \frac{\langle \tilde{x}, \tilde{y} \rangle}{||\tilde{x}||\ ||\tilde{y}||} \leq 1 \\ -1 \leq corr(x, y) \leq 1 \end{aligned}$$

Joy Yang, October 12, 2018

3 Linear Regression

- Model: $Y_{n\times 1} = X_{n\times p}\beta_{p\times 1}$

$$X = \begin{pmatrix} - & x_1^T & - \\ & \cdots & \\ - & x_n^T & - \end{pmatrix}$$

- How do we find a reasonable value for β
- Idea: minimize the squared error

$$\begin{aligned}\hat{\beta} &= argmin_\beta \sum_{i=1}^{n} (y_i - x_i^T\beta)^2 \\ &= argmin_\beta (Y - X\beta)^T(Y - X\beta)\end{aligned}$$

- **Property:** $\nabla_x(u^T v) = u^T(\nabla_x v) + v^T(\nabla_x u)$
- **Question:** Find β

$$\begin{aligned}0 &\doteq \nabla_\beta\ (Y - X\beta)^T(Y - X\beta) \\ 0 &= -2\ (Y - X\beta)^T(X) \\ X^T X\beta &= X^T Y \\ \hat{\beta} &= (X^T X)^{-1}X^T Y\end{aligned}$$

And our predicted values

$$\begin{aligned}\hat{Y} &= X\hat{\beta} \\ \hat{Y} &= X(X^T X)^{-1}X^T Y\end{aligned}$$

4 Normal Errors and MLE

- Let's rewrite our model to account for noise:

  $y_i = x_i\beta + \epsilon_i$, $\epsilon_i \sim N(0, 1)$

- It turns out that this implies Y is a multivariate normal:

  $Y \sim N(X\beta, I)$

- **Definition:** A multivariate normal random variable $z \sim N(\mu, \Sigma)$ has a pdf:

$$f(z; \mu, \Sigma) = det(2\pi\Sigma)^{-\frac{1}{2}} exp(-\frac{1}{2}(z-\mu)^T \Sigma^{-1}(z-\mu))$$

- **Question:** Find the maximum likelihood estimator (MLE) for β

The pdf for Y is

$$f(Y; X\beta, I) = det(2\pi I)^{-\frac{1}{2}} exp(-\frac{1}{2}(Y - X\beta)^T(Y - X\beta))$$

So the log likelihood function for β is

$$l(\beta) = log(\text{some constant wrt } \beta) - \frac{1}{2}(Y - X\beta)^T(Y - X\beta)$$

To find the MLE for β, we need to maximize the function above. And simplifying:

$$\hat{\beta}_{MLE} = argmax_\beta - (Y - X\beta)^T(Y - X\beta)$$

Notice that this is equivalent to taking the *argmin* of the sum of squared error.

$$\hat{\beta}_{SSE} = argmin_\beta(Y - X\beta)^T(Y - X\beta)$$

5 Vector Spaces

- X is $n \times p$ matrix.

$$X = \begin{pmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_p \\ | & | & & | \end{pmatrix}$$

- Column space: $C(X) = \{w : w = Xc\}$
- Null space: $N(X) = \{w : Xw = 0\}$
- Orthogonal complement: $C(X)^{\perp} = \{w : w^T X = 0\}$
- **Question:** Show that $\hat{Y}$ (our vector of fitted values) lives in $C(X)$, and $Y - \hat{Y}$ (our vector of residuals) lives in $C(X)^{\perp}$.

If $\hat{Y}$ is in $C(x)$, then we should be able to find a vector c such that $\hat{Y} = Xc$. This is easy, because we defined $\hat{Y} = X\hat{\beta}$.

If $Y - \hat{Y}$ is in $C(X)^{\perp}$ then $(Y - \hat{Y})X = 0$. Remember that in finding β in Section 2, we'd taken the gradient of the sum of squared errors and set that to 0. This gave us the condition that $2\ (Y - X\hat{\beta})^T(-X) = 0$, which is equivalent to the condition that $(Y - \hat{Y})X = 0$.

6 Eigenvectors and Eigenvalues

- For the matrix A, v is an **eigenvector** corresponding to the **eigenvalue** λ if

  $$Av = \lambda v$$

  Here λ is a scalar

- The intuition is that A only stretches v and does not rotate the vector

- Here's an example:

  $$A = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

  The eigenvectors of A are $\left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right)$ with corresponding eigenvalues 1.5 and 0.5.

- **Spectral Decomposition:** We can decompose any real, symmetric matrix A as

  $$A = Q\Lambda Q^T$$

  Where the columns of U are the **orthogonal eigenvectors** of A and Λ contains the **eigenvalues** of A (everything is real). Q is an orthogonal matrix, which means it is a square matrix and that the columns are orthogonal and have norm 1.

- Back to the previous example, we can rescale our eigenvectors and eigenvalues to get Q and λ. The current norm of the two vectors is $\sqrt{2}$ so

  $$Q = \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix}$$

  $$\Lambda = \begin{pmatrix} \frac{3\sqrt{2}}{4} & 0 \\ 0 & \frac{\sqrt{2}}{4} \end{pmatrix}$$

- Why do we care about Spectral Decomposition?
  - Spectral decomposition gives us a really nice way of getting pseudoinverses, as we can just take the inverse of Λ. Taking the inverse of Λ is easy, because it's a diagonal matrix, so the inverse is diagonal with entries $1/\lambda_{ii}$.
  - It also gives us a nice way of taking matrix square roots (we can just take the square root of λ).
  - There's a nice interpretation of the eigenvectors as the principal axes of variation. This will come into play again when we talk about PCA next week.