CPSC 340:
Machine Learning and Data Mining

The Normal Equations
Fall 2017
Gradient and Critical Points in d-Dimensions

• Generalizing “set the derivative to 0 and solve” in d-dimensions:
  – Find ‘w’ where the gradient vector equals the zero vector.
• Gradient is vector with partial derivative ‘j’ in position ‘j’:

$$\nabla f(w) = \begin{bmatrix} 2f \\ 2w_1 \\ 2f \\ 2w_2 \\ \vdots \\ 2f \\ 2w_d \end{bmatrix}$$

Tangent slope is 0 in every direction at minimizers.
Gradient and Critical Points in d-Dimensions

• Generalizing “set the derivative to 0 and solve” in d-dimensions:
  – Find ‘w’ where the gradient vector equals the zero vector.
• Gradient is vector with partial derivative ‘j’ in position ‘j’:

\[ \nabla f(w) = \begin{bmatrix} 2f \\ 2w_1 \\ 2f \\ 2w_2 \\ \vdots \\ 2f \\ 2w_d \end{bmatrix} \]

For linear least squares:

\[ \nabla f(w) = \begin{bmatrix} \sum_{i=1}^{n} (w^\top x_i - y_i) x_{i1} \\ \sum_{i=1}^{n} (w^\top x_i - y_i) x_{i2} \\ \vdots \\ \sum_{i=1}^{n} (w^\top x_i - y_i) x_{id} \end{bmatrix} \]

Claims for linear least squares:

1. Finding a ‘w’ where \( \nabla f(w)=0 \) can be done by solving a System of linear equations.
2. All ‘w’ where \( \nabla f(w)=0 \) are minimizers.
Least Squares in d-Dimensions

- The linear least squares model in d-dimensions minimizes:
  \[ f(w) = \frac{1}{2} \sum_{i=1}^{n} (w^T x_i - y_i)^2 \]

- Computing the partial derivative:
  \[ \frac{\partial}{\partial w_j} \left[ \frac{1}{2} \sum_{i=1}^{n} (w^T x_i - y_i)^2 \right] = \frac{1}{2} \sum_{i=1}^{n} \frac{\partial}{\partial w_j} \left[ (w^T x_i - y_i)^2 \right] 
  = \frac{1}{2} \sum_{i=1}^{n} 2 (w^T x_i - y_i) \frac{\partial}{\partial w_j} [w^T x_i] 
  = \sum_{i=1}^{n} (w^T x_i - y_i) x_{il} \]

Problem: I can't just set to 0 and solve because it depends on \( w_2, w_3, \ldots, w_d \)
Matrix/Norm Notation (MEMORIZE/STUDY THIS)

• To solve the $d$-dimensional least squares, we use matrix notation:
  – We use ‘$y$’ as an “$n$ times 1” vector containing target ‘$y_i$’ in position ‘$i$’.
  – We use ‘$x_i$’ as a “$d$ times 1” vector containing features ‘$j$’ of example ‘$i$’.
    - We’re now going to be careful to make sure these are column vectors.
  – So ‘$X$’ is a matrix with the $x_i^T$ in row ‘$i$’.

\[
\begin{align*}
  y &= \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, &
  x_i &= \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{bmatrix}, &
  X &= \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}
\end{align*}
\]
Matrix/Norm Notation (MEMORIZE/STUDY THIS)

- To solve the d-dimensional least squares, we use matrix notation:
  - Our prediction for example ‘i’ is given by scalar $w^T x_i$.
  - The matrix-vector product $Xw$ gives predictions for all ‘i’ (n times 1 vector).

\[
\begin{align*}
  w^T x_i &= \sum_{j=1}^{d} w_j x_{ij} \\
  &= w_1 x_{i1} + w_2 x_{i2} + \ldots + w_d x_{id}
\end{align*}
\]

Also, because $w^T x_i$ is a scalar, we have $w^T x_i = x_i^T w$.
(e.g., $[5]^T = [5]$)

\[
Xw = \begin{bmatrix}
  x_{11} & x_{12} & \cdots & x_{1d} \\
  x_{21} & x_{22} & \cdots & x_{2d} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{n1} & x_{n2} & \cdots & x_{nd}
\end{bmatrix} \begin{bmatrix}
  w_1 \\
  w_2 \\
  \vdots \\
  w_d
\end{bmatrix} = \begin{bmatrix}
  x_{11}w_1 + x_{12}w_2 + \cdots + x_{1d}w_d \\
  x_{21}w_1 + x_{22}w_2 + \cdots + x_{2d}w_d \\
  \vdots \\
  x_{n1}w_1 + x_{n2}w_2 + \cdots + x_{nd}w_d
\end{bmatrix}
\]

\[
= \begin{bmatrix}
  \sum_{j=1}^{d} x_{1j}w_j \\
  \sum_{j=1}^{d} x_{2j}w_j \\
  \vdots \\
  \sum_{j=1}^{d} x_{nj}w_j
\end{bmatrix} = \begin{bmatrix}
  x_1^T w \\
  x_2^T w \\
  \vdots \\
  x_n^T w
\end{bmatrix} = \begin{bmatrix}
  w_1^T x_1 \\
  w_2^T x_2 \\
  \vdots \\
  w_d^T x_d
\end{bmatrix}
\]

Prediction for example ‘i’ is in column ‘i’.
Matrix/Norm Notation (MEMORIZE/STUDY THIS)

• To solve the d-dimensional least squares, we use **matrix notation**:
  – Our **prediction for example ‘i’** is given by scalar $w^T x_i$.
  – The **matrix-vector product $Xw$** gives predictions for all ‘i’ (n times 1 vector).
  – The **residual vector $r$** gives $w^T x_i$ minus $y_i$ for all ‘i’ (n times 1 vector).
  – Least squares can be written as the squared L2-norm of the residual.

\[
\begin{align*}
\mathbf{r} &= \begin{bmatrix}
  w^\top x_1 - y_1 \\
  w^\top x_2 - y_2 \\
  \vdots \\
  w^\top x_n - y_n \\
\end{bmatrix}
= \begin{bmatrix}
  w^\top x_1 \\
  w^\top x_2 \\
  \vdots \\
  w^\top x_n \\
\end{bmatrix}
- \begin{bmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_n \\
\end{bmatrix}
= Xw - y
\end{align*}
\]

\[
\sum_{i=1}^{n} (w^\top x_i - y_i)^2 = \sum_{i=1}^{n} (r_i)^2 \\
= \sum_{i=1}^{n} r_i r_i \\
= r^\top r \\
= \|r\|^2 = \|Xw - y\|^2
\]
Matrix Algebra Review (MEMORIZE/STUDY THIS)

• Review of linear algebra operations we’ll use:
  – If ‘a’ and ‘b’ be vectors, and ‘A’ and ‘B’ be matrices then:

\[
\begin{align*}
\mathbf{a}^\top \mathbf{b} &= \mathbf{b}^\top \mathbf{a} \\
\|\mathbf{a}\|_2 &= \mathbf{a}^\top \mathbf{a} \\
(\mathbf{A} + \mathbf{B})^\top &= \mathbf{A}^\top + \mathbf{B}^\top \\
(\mathbf{AB})^\top &= \mathbf{B}^\top \mathbf{A}^\top \\
(\mathbf{A} + \mathbf{B})(\mathbf{A} + \mathbf{B}) &= \mathbf{AA} + \mathbf{BA} + \mathbf{AB} + \mathbf{BB} \\
\mathbf{a}^\top \mathbf{A} \mathbf{b} &= \mathbf{b}^\top \mathbf{A}^\top \mathbf{a}
\end{align*}
\]

Sanity check: ALWAYS CHECK THAT DIMENSIONS MATCH (if not, you did something wrong)
Linear Least Squares

Want \( \mathbf{w} \) that minimizes
\[
f(w) = \frac{1}{2} \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i - y_i)^2 = \frac{1}{2} \| \mathbf{Xw} - \mathbf{y} \|_2^2
\]
\[
= \frac{1}{2} (\mathbf{Xw} - \mathbf{y})^T (\mathbf{Xw} - \mathbf{y})
\]
\[
= \frac{1}{2} (\mathbf{w}^T \mathbf{X}^T - \mathbf{y}^T) (\mathbf{Xw} - \mathbf{y})
\]
\[
= \frac{1}{2} \left( \mathbf{w}^T \mathbf{X}^T (\mathbf{Xw} - \mathbf{y}) - \mathbf{y}^T (\mathbf{Xw} - \mathbf{y}) \right)
\]
\[
= \frac{1}{2} \left( \mathbf{w}^T \mathbf{X}^T \mathbf{Xw} - \mathbf{w}^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{Xw} + \mathbf{y}^T \mathbf{y} \right)
\]
\[
= \frac{1}{2} \mathbf{w}^T \mathbf{X}^T \mathbf{Xw} - \frac{1}{2} \mathbf{w}^T \mathbf{X}^T \mathbf{y} + \frac{1}{2} \mathbf{y}^T \mathbf{y}
\]

Sanity check: all of these are scalars.
Linear and Quadratic Gradients

• We’ve written as a $d$-dimensional quadratic:

$$f(w) = \sum_{i=1}^{D} (w^T x_i - y_i)^2 = \frac{1}{2} \|Xw - y\|^2 = \frac{1}{2} w^T X^T Xw - w^T X^T y + \frac{1}{2} y^T y$$

$$= \frac{1}{2} w^T A w + w^T b + c$$

• How do we compute gradient?

Let’s first do it with $d=1$:

$$f(w) = \frac{1}{2} w^2 + wb + c$$

$$f'(w) = aw + b + 0$$

Here are the generalizations to $d$ dimensions:

$$\nabla [c] = 0 \text{ (zero vector)}$$

$$\nabla [w^T b] = b$$

$$\nabla [w^T A w] = A w \text{ (if } A \text{ is symmetric)}$$

Full derivations are on webpage in notes on linear and quadratic gradients.
Linear and Quadratic Gradients

• We’ve written the least squares objective as a quadratic function:

\[
f(w) = \frac{1}{2} \sum_{i=1}^{n} (w^T x_i - y_i)^2 = \frac{1}{2} \| X_w - y \|^2 = \frac{1}{2} w^T X^T X w - w^T X^T y + \frac{1}{2} y^T y
\]

\[
= \frac{1}{2} w^T A w + w^T b + c
\]

• Gradient is given by:

\[
\nabla f(w) = A w + b + 0
\]

• Using definitions of ‘A’ and ‘b’:

\[
= X^T X w + X^T y
\]

Sanity check: these are both \( d \times 1 \) vectors.
Normal Equations

• Set gradient equal to zero to find the least squares “critical points”:
  \[ X^\top X_w - X^\top y = 0 \]

• We now move terms not involving ‘w’ to the other side:
  \[ X^\top X_w = X^\top y \]

• This is a set of ‘d’ linear equations called the normal equations.
  – This a linear system like “Ax = b” from Math 152.
    • You can use Gaussian elimination to solve for ‘w’.
  – In Julia, the “\” command can be used to solve linear systems:

 Train: \[ w = (X^\top X) \backslash (X^\top y) \]

 Predict: \[ \hat{y} = X_{\text{test}} \ast w \]
Incorrect Solutions to Least Squares Problem

The least squares objective is \( f(w) = \frac{1}{2} \| Xw - y \|^2 \)

The minimizers of this objective are solutions to the linear system:
\[
X^T X w = X^T y
\]

The following are not the solutions to the least squares problem:

\[
w = (X^T X)^{-1}(X^T y) \quad \text{(only true if } X^T X \text{ is invertible)}
\]

\[
w X^T X = X^T y \quad \text{(matrix multiplication is not commutative, dimensions don’t even match)}
\]

\[
w = \frac{X^T y}{X^T X} \quad \text{(you cannot divide by a matrix)}
\]
Least Squares Issues

• Issues with least squares model:
  – Solution might **not** be unique.
  – It is **sensitive** to outliers.
  – It always uses all features.
  – Data can might so big we can’t store $X^T X$.
  – It might predict outside range of $y_i$ values.
  – It assumes a **linear** relationship between $x_i$ and $y_i$.

$X$ is $n \times d$

so $X^T$ is $d \times n$

and $X^T X$ is $d \times d$.

Costs $O(nd^2)$ to calculate:
  – Each of the $O(d^2)$ elements
    is an inner product between length $n$ vectors.
Non-Uniqueness: Colinearity

• Imagine have two features that are identical for all examples.
• Then these features are called **collinear**.
• I can increase weight on one feature, and decrease it on the other, **without changing predictions**.
• Thus the solution is not unique.

• But, any ‘w’ where $\nabla f(w) = 0$ is a global optimum, due to **convexity**.
Convex Functions

• Is finding a ‘w’ with $\nabla f(w) = 0$ good enough?
  – Yes, for convex functions.

• A function is convex if the area above the function is a convex set.
  – All values between any two points above function stay above function.
Convex Functions

• All ‘w’ with $\nabla f(w) = 0$ for convex functions are global minima.

Proof by contradiction:

Consider a local minimum. If this is not global minimum, there must a smaller value.

By convexity we can move along line to global minimum and decrease objective.

– Normal equations finds a global minimum because of convexity.
How do we know if a function is convex?

- Some useful tricks for showing a function is convex:
  - 1-variable, twice-differentiable function is convex iff \( f''(w) \geq 0 \) for all ‘w’.

Consider \( f(w) = \frac{1}{2} aw^2 \) for \( a > 0 \). We have \( f'(w) = aw \) and \( f''(w) = a > 0 \) by assumption.

Consider \( f(w) = e^w \). We have \( f'(w) = e^w \) and \( f''(w) = e^w > 0 \) by definition of exponential function.
How do we know if a function is convex?

• Some useful tricks for showing a function is convex:
  – 1-variable, twice-differentiable function is convex iff \( f''(w) \geq 0 \) for all ‘\( w \)’.
  – A convex function multiplied by non-negative constant is convex.

We showed that \( f(w) = e^w \) is convex, so \( f(w) = 10e^w \) is convex.
How do we know if a function is convex?

- Some useful tricks for showing a function is convex:
  - 1-variable, twice-differentiable function is convex iff $f''(w) \geq 0$ for all $w$.
  - A convex function multiplied by non-negative constant is convex.
  - Norms and squared norms are convex.

\[\|w\|, \|w\|^2, \|w\|_1, \|w\|_\infty, \|w\|^2_2\] and so on are all convex.
How do we know if a function is convex?

• Some useful tricks for showing a function is convex:
  – 1-variable, twice-differentiable function is convex iff $f''(w) \geq 0$ for all ‘w’.
  – A convex function multiplied by non-negative constant is convex.
  – Norms and squared norms are convex.
  – The sum of convex functions is a convex function.

$$f(x) = 10e^w + \frac{1}{2} \|w\|^2 \quad \text{is convex}$$

From earlier constant norm squared
How do we know if a function is convex?

• Some useful tricks for showing a function is convex:
  – 1-variable, twice-differentiable function is convex iff \( f''(w) \geq 0 \) for all ‘\( w \)’.
  – A convex function multiplied by non-negative constant is convex.
  – Norms and squared norms are convex.
  – The sum of convex functions is a convex function.
  – The max of convex functions is a convex function.

\[
 f(w) = \max \sum \left\{ 2w, w^2 \right\} \text{ is convex.}
\]
How do we know if a function is convex?

• Some useful tricks for showing a function is convex:
  – 1-variable, twice-differentiable function is convex iff $f''(w) \geq 0$ for all ‘$w$’.
  – A convex function multiplied by non-negative constant is convex.
  – Norms and squared norms are convex.
  – The sum of convex functions is a convex function.
  – The max of convex functions is a convex function.
  – Composition of a convex function and a linear function is convex.

\[
\text{If } f \text{ is convex the } f(Xw - y) \text{ is convex.}
\]
How do we know if a function is convex?

• Some useful tricks for showing a function is convex:
  – 1-variable, twice-differentiable function is convex iff \( f''(w) \geq 0 \) for all \( w \).
  – A convex function multiplied by non-negative constant is convex.
  – Norms and squared norms are convex.
  – The sum of convex functions is a convex function.
  – The max of convex functions is a convex function.
  – Composition of a convex function and a linear function is convex.

• But: not true that composition of convex with convex is convex:
  Even if \( f \) is convex and \( g \) is convex, \( f(g(w)) \) might not be convex.
  E.g. \( x^2 \) is convex and \(-\log(x)\) is convex but \(-\log(x^2)\) is not convex.
Example: Convexity of Linear Regression

• Consider linear regression objective with squared error:

\[ f(w) = \|Xw - y\|^2 \]

• We can use that this is a **convex function composed with linear**:

Let \( g(r) = \|r\|^3 \), which is **convex** because it's a squared norm.

Then \( f(w) = g(Xw - y) \), which is **convex** because it's a convex function composed with the linear function \( h(w) = Xw - y \).
Summary

• Normal equations: solution of least squares as a linear system.
  – Solve \((X^TX)w = (X^ty)\).

• Solution might not be unique because of collinearity.

• But any solution is optimal because of convexity.

• Convex functions:
  – Set of functions with property that \(\nabla f(w) = 0\) implies ‘w’ is a global min.
  – Can (usually) be identified using a few simple rules.

• Next time: overview of numerical optimization concepts.
Convexity, min, and argmin

• If a function is convex, then all stationary points are global optima.

• However, convex functions don’t necessarily have stationary points:
  – For example, $f(x) = a*x$, $f(x) = \exp(x)$, etc.

• Also, more than one ‘$x$’ can achieve the global optimum:
  – For example, $f(x) = c$ is minimized by any ‘$x$’.
• **Householder notation**: set of (fairly-logical) conventions for math.

- Use **Greek letters** for scalars: \( \alpha = 1, \beta = 3.5, \gamma = \pi \)
- Use **first/last lowercase** letters for vectors: \( \mathbf{w} = \begin{bmatrix} 0.1 \end{bmatrix}, \mathbf{x} = \begin{bmatrix} 1 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} -1 \end{bmatrix}, \mathbf{a} = \begin{bmatrix} 1 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 0.5 \end{bmatrix} \)
  - Assumed to be **column-vectors**.
- Use **first/last uppercase** letters for matrices: \( X, Y, W, A, B \)

- Indices use \( i, j, k \).
- Sizes use \( m, n, d, p \) and \( k \) is obvious from context.
- Sets use \( S, T, U, V \).
- Functions use \( f, g, h \).

When I write \( x_i \), I mean "grab row \( i \) of \( X \) and make a **column-vector** with its values."
Bonus Slide: Householder(-ish) Notation

- **Householder notation**: set of (fairly-logical) conventions for math:

Our ultimate least squares notation:

\[ f(w) = \frac{1}{2} \|Xw - y\|^2 \]

But if we agree on notation we can quickly understand:

\[ g(x) = \frac{1}{2} \|Ax - b\|^2 \]

If we use random notation we get things like:

\[ H(\beta) = \frac{1}{2} \|R\beta - p\|^2 \]

Is this the same model?
When does least squares have a unique solution?

• We said that least squares solution is not unique if we have repeated columns.
• But there are other ways it could be non-unique:
  – One column is a scaled version of another column.
  – One column could be the sum of 2 other columns.
  – One column could be three times one column minus four times another.
• Least squares solution is unique if and only if all columns of $X$ are “linearly independent”.
  – No column can be written as a “linear combination” of the others.
  – Many equivalent conditions (see Strang’s linear algebra book):
    • $X$ has “full column rank”, $X^TX$ is invertible, $X^TX$ has non-zero eigenvalues, $\det(X^TX) > 0$.
    – Note that we cannot have independent columns if $d > n$. 