Machine learning lecture slides

Regression I: Linear regression

Outline

- Statistical model for regression
- College GPA example
- Ordinary least squares for linear regression
- The expected mean squared error

Figure 1: Galton board

Real-valued predictions

- Example: Galton board
- Physical model: hard
- Statistical model: final position of ball is random
  - *Normal (Gaussian) distribution* with mean μ and variance σ^2

$$\frac{1}{\sqrt{2\pi\sigma^2}}$$

- Goal: predict final position accurately, me
(also called *squared error*)

$$(\text{prediction} - \text{outcome})^2$$

  - Outcome is random, so look at *expected squared loss* (also called *mean squared error*)

Optimal prediction for mean squared error

- Predict $\hat{y} \in \mathbb{R}$; true final position is Y (random variable) with *mean* $\mathbb{E}(Y) = \mu$ and *variance* $\mathrm{var}(Y) = \mathbb{E}[(Y - \mathbb{E}(Y))^2] = \sigma^2$.
- Squared error is $(\hat{y} - Y)^2$.
- *Bias-variance decomposition*:

$$\mathbb{E}[(\hat{y} - Y)^2]$$

$$y - \mu \qquad \sigma \ .$$

- This is true for any random variable assumption.
- So optimal prediction is $\hat{y} = \mu$.
- When parameters are unknown, can estimate from related data, ...
- Can also do an analysis of a plug-in prediction ...

Statistical model for regression

- Setting is same as for classification except:
  - Label is real number, rather than $\{0, 1\}$ or $\{1, 2, \ldots, K\}$
  - Care about squared loss, rather than whether prediction is correct
  - *Mean squared error* of f:

Optimal prediction function for regression

- If (X, Y) is random test example, then *optimal prediction function* is

$$f^\star(x) = \mathbb{E}[Y \mid X = x]$$

Test MSE (1)

- Just like in classification, we can use test data to estimate $\mathrm{mse}(\hat{f})$ for a function $\hat{f}$ that depends only on training data.
- IID model: $(X_1, Y_1), \ldots, (X_n, Y_n), (X'_1, Y'_1), \ldots, (X'_m, Y'_m), (X, Y)$ are iid

- Predictor $\hat{f}$ is based only on training exa
- Hence, **test examples are independe** important)
- We would like to estimate $\mathrm{mse}(\hat{f}$

Test MSE (2)

- Test MSE $\mathrm{mse}(\hat{f}, T) = \frac{1}{m}\sum_{i=1}^{m}(\hat{f}(X_i') \neq Y_i')^2$
  - By law of large numbers, $\mathrm{mse}(\hat{f}, T) \to \mathrm{mse}(\hat{f})$ as $m \to \infty$

Example: College GPA

- Data from 750 Dartmouth students' College GPA
  - Mean: 2.46
  - Standard deviation: 0.746
- Assume this data is iid sample from the population of Dartmouth students (false)
-

Figure 2: Histogram of College GPA

Predicting College GPA from HS GPA (1)

- Students represented in data have High School (HS) GPA
  - Maybe HS GPA is predictive of College GPA?
- Data: $S := ((x_1, y_1), \ldots, (x_n, y_n))$
  - x_i is HS GPA of i-th student

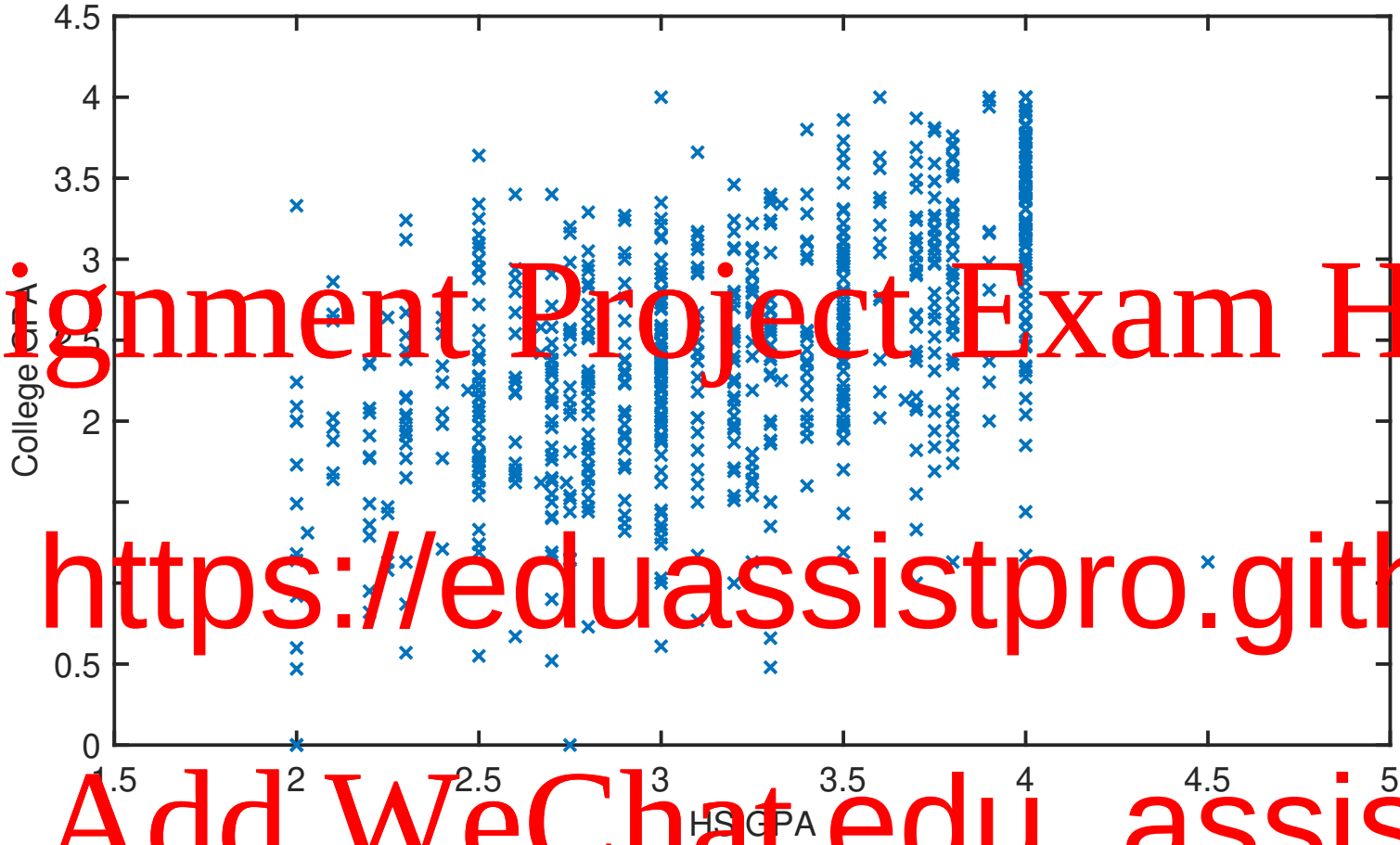

Figure 3: Plot of College GPA vs HS G

Predicting College GPA from HS GPA (2)

- First attempt:
  - Define intervals of possible HS GPAs:

$$(0.00, 0.25], \quad (0.25, 0.50], \quad (0.50, 0.75], \quad \ldots$$

$$\hat{f}(x) := \begin{cases} \hat{\mu}_{(0.25,0.50]} \\ \hat{\mu}_{(0.50,0.75]} \\ \end{cases}$$

  - (What to do about an interval I that contains no student's HS GPA?)

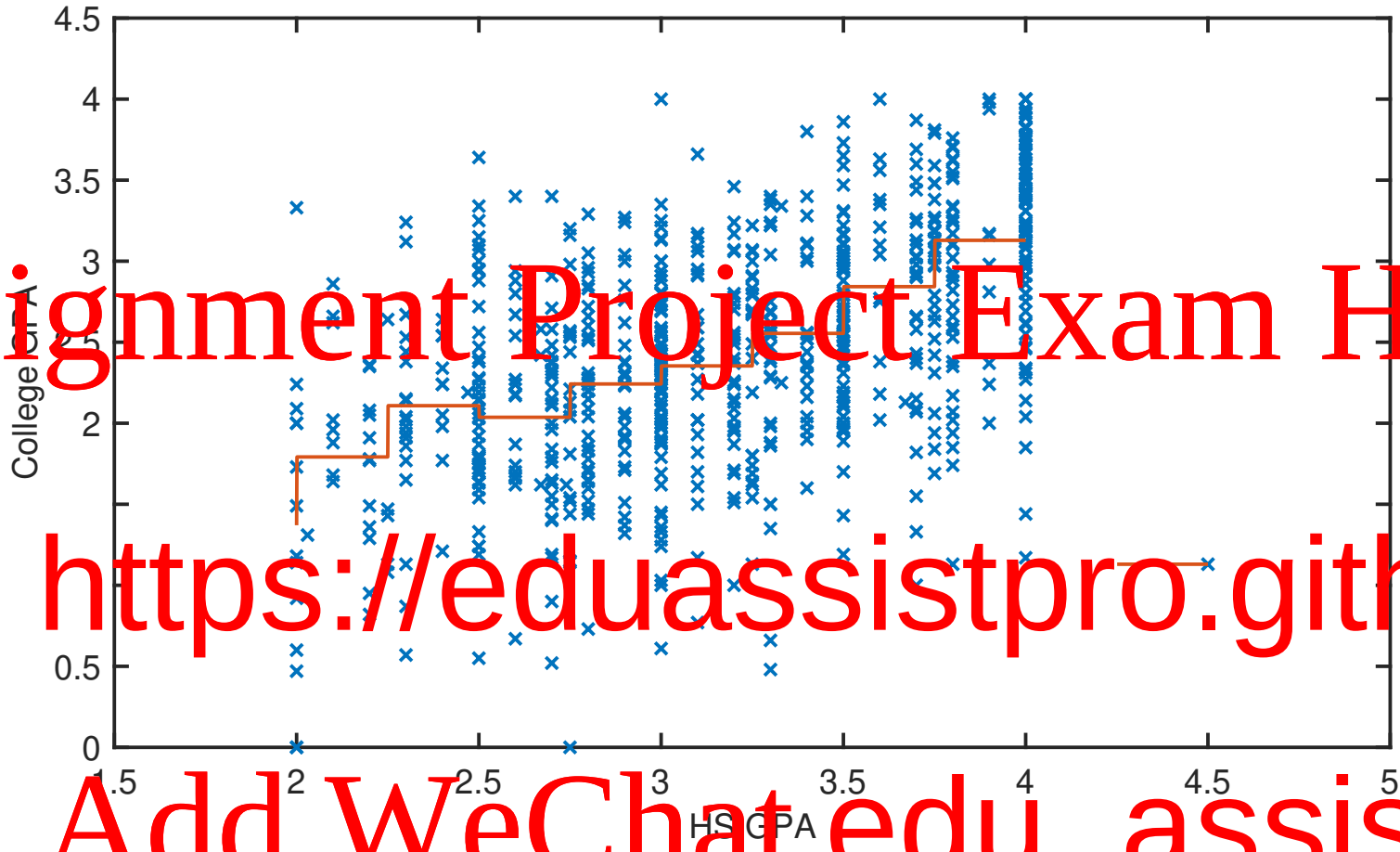

Figure 4: Plot of mean College GPA vs binne

Predicting College GPA from HS GPA (3)

- Define

$$\text{mse}(f, S) := \frac{1}{|S|} \sum_{x,y \in S} (f(x) - y)^2,$$

the mean squared error of predictions made by f on examples

$$\text{mse}(\hat{f}, S) = 0.376$$

$$\sqrt{\text{mse}(\hat{f}, S)} = 0.613 < 0.746 \ (\text{std. dev. of } y_i\text{'s})$$

- Piece-wise constant function $\hat{f}$ is an improvement over the constant function (i.e., just predicting the mean 2.46 for all x)!

Predicting College GPA from HS GPA (4)

- But $\hat{f}$ has some quirks.
- E.g., those with HS GPA between 2.50 and 2.75 are predicted to have a lower College GPA than those with HS GPA between 2.25 and 2.50.
- E.g., something unusual with the student who has HS GPA of

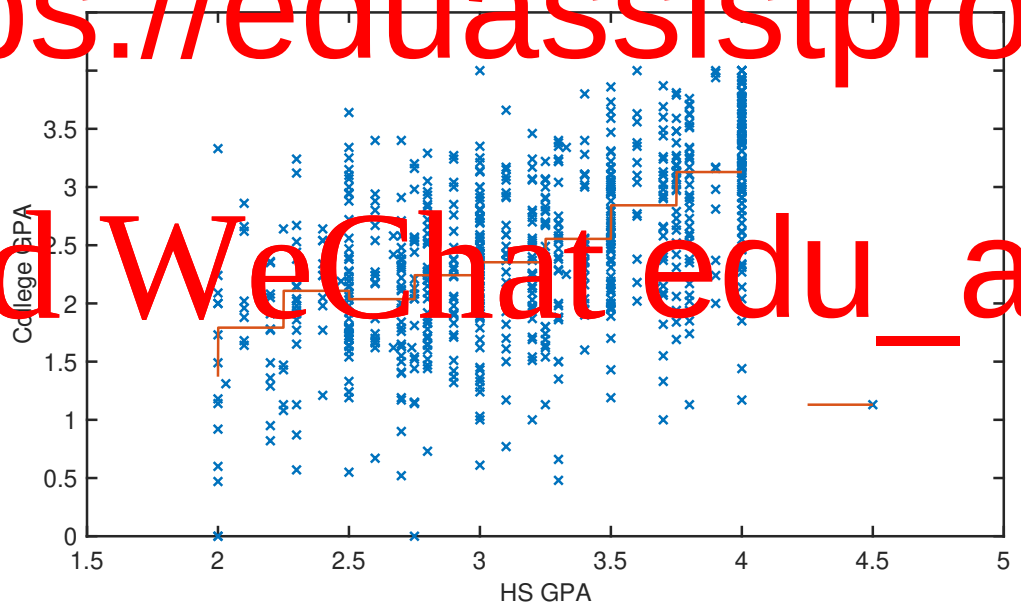

Figure 5: Plot of mean College GPA vs binned HS GPA

Least squares linear regression (1)

- Suppose we'd like to only consider functions with a specific functional form, e.g., a linear function:

$$f(x) = mx + \theta$$

m

prediction of College GPA.

Least squares linear regression (2)

- What is the linear function with smallest MSE on $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R} \times \mathbb{R}$? This is the problem of *least squares linear regression*
  - Find $(m, \theta) \in \mathbb{R}^2$ to minimize

- Also called *ordinary least squares*

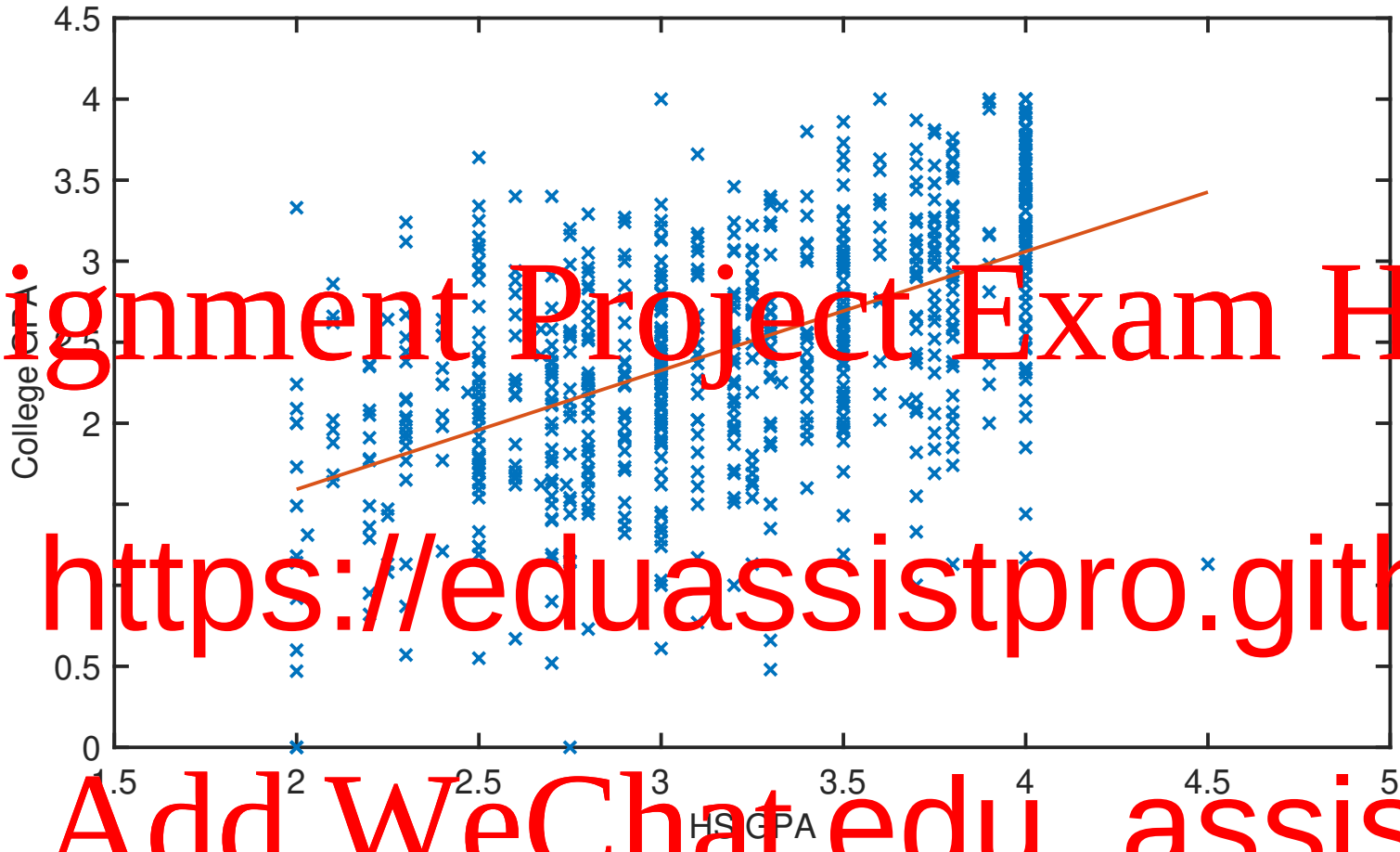

Figure 6: Plot of least squares linear regre

Computing OLS (1)

- Derivatives equal zero conditions (*normal equations*):

$$\frac{\partial}{\partial \theta}\left\{\frac{1}{n}\sum_{i=1}^{n}(mx_i+\theta-y_i)^2\right\} = \frac{2}{n}\sum_{i=1}^{n}(mx_i+\theta-y_i) = 0$$

$$\frac{\partial}{\partial m}\left\{\frac{1}{n}\sum_{i=1}^{n}(mx_i+\theta-y_i)^2\right\} = \frac{2}{n}\sum_{i=1}^{n}(mx_i+\theta-y_i)x_i = 0.$$

- Define

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{xy} := \frac{1}{n}\sum_{i=1}^{n} x_i y_i, \qquad \overline{y} := \frac{1}{n}\sum_{i=1}^{n} y_i,$$

so system can be re-written as

$$\overline{x}m + \theta = \overline{y}$$

$$\overline{x^2}m + \overline{x}\theta = \overline{xy}.$$

Computing OLS (2)

- Write in matrix notation:

$$\begin{bmatrix} \overline{x^2} & \overline{x} \\ \overline{x} & 1 \end{bmatrix} \begin{bmatrix} \hat{m} \\ \hat{\theta} \end{bmatrix} = \begin{bmatrix} \overline{xy} \\ \overline{y} \end{bmatrix}.$$

- Solution: $(\hat{m}, \hat{\theta}) \in \mathbb{R}^2$ given by

$$\hat{m} := \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2}, \quad \hat{\theta} := \overline{y} - \hat{m}\,\overline{x}$$

Computing OLS (3)

- Catch: The above solution only makes sense if $\overline{x^2} - \overline{x}^2 \neq 0$, i.e., the variance of the x_i's is non-zero.

- If $\overline{x^2} - \overline{x}^2 = 0$, then the matrix defining t equations is singular.

Computing OLS (4)

- In general, "derivative equals zero" is only a necessary condition for a solution to be optimal; not necessarily a sufficient condition!

- **Theorem** Every solution to the normal eq optimal solution to the least squares linear r

Decomposition of expected MSE (1)

- Two different functions of HS GPA for predicting College GPA.
  - What makes them different?
  - We care about prediction of College GPA for student we haven't seen before based on their HS GPA.

$$
\begin{aligned}
\mathbb{E}[\mathrm{mse}(\hat{f})] \\
&= \mathbb{E}\left[\mathbb{E}[(\hat{f}(X) - Y)^2 \mid \hat{f}]\right] \\
&= \mathbb{E} \\
&= \mathbb{E} \\
&= \mathbb{E}\ \ \mathrm{var}(Y \mid X) + \mathbb{E}[(f(X) - \mathbb{E}[Y \mid \quad {}^2 \\
&= \mathbb{E}\left[\mathrm{var}(Y \mid X) + \mathrm{var}(\hat{f}(X) \mid X) + (\qquad)^2\right] \\
&= \underbrace{\mathbb{E}\left[\mathrm{var}(Y \mid X)\right]}_{\text{unavoidable error}} + \underbrace{\mathbb{E}\left[\mathrm{var}(\hat{f}(X) \mid X)\right.}_{\text{variability of } \hat{f}} \underbrace{\left. f\ X\ \mid X\ -\ [Y \mid X])^2\right]}_{\text{approximation error of } \hat{f}}
\end{aligned}
$$

Decomposition of expected MSE (3)

- First term is quantifies inherent unpredictability of Y (even after seeing X)
- Second term measures the "variability" of $\hat{f}$ due to the random nature of training data. Depends on:

- Third term quantifies how well a function pr fitting procedure can approximate the reg after removing the "variability" of

Multivariate linear regression (1)

- For Dartmouth data, also have SAT Score for all students.
  - Can we use both *predictor variables* (HS GPA and SAT Score) to get an even better prediction of College GPA?
  - Binning approach: instead of a 1-D grid (intervals), consider a 2-D grid (squares).

for some $(m_1, m_2) \in \mathbb{R}^2$ and θ

Multivariate linear regression (2)

- The general case: a (homogeneous) linear function $f : \mathbb{R}^d \to \mathbb{R}$ of the form

$$f(x) = x^T w$$

for some $w \quad \mathbb{R}^d$.

Then

Multivariate ordinary least squares (1)

- What is the linear function with smallest MSE on $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$?
  - Find $v \in \mathbb{R}^d$ to minimize

$$\frac{1}{n} \sum_{} \quad \mathsf{T} \quad 2$$

Multivariate ordinary least squares (2)

- In matrix notation:

$$\widehat{R}(w) := \frac{1}{n}\|Aw - b\|_2^2$$

where

$$A := \frac{1}{\sqrt{n}} \underbrace{\begin{bmatrix} x_1^\mathsf{T} \\ \vdots \\ x_n^\mathsf{T} \end{bmatrix}}_{n} \in \mathbb{R}^{n \times d} \qquad b := \frac{1}{\sqrt{n}} \underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{n} \in \mathbb{R}^n$$

- If we put vector $v \in \mathbb{R}^d$ in the con
  it is treated as a column vector by default!
  - If we want a row vector, we write
- Therefore

$$Aw - b = \frac{1}{\sqrt{n}} \begin{bmatrix} x_1^\mathsf{T} w - y_1 \\ \vdots \\ x_n^\mathsf{T} w - y_n \end{bmatrix}$$

Figure 7: Geometric picture of least squares linear regression

Multivariate normal equations (1)

- Like the one-dimensional case, optimal solutions are characterized by a system of linear equations (the "derivatives equal zero" conditions) called the *normal equations*:

$$\frac{\partial \quad (w)}{}$$

$$\overline{\partial w_d}$$

which is equivalent to

$$A^\mathsf{T} A w =$$

Multivariate normal equations (2)

- If $A^\mathsf{T}A$ is non-singular (i.e., invertible), then there is a unique solution given by

$$\hat{w} := (A^\mathsf{T}A)^{-1}A^\mathsf{T}b.$$

-

- **Theorem**: Every solution to the normal eq optimal solution to the least squares linear r

Algorithm for least squares linear regression

- How to solve least squares linear regression problem?
  - Just solve the normal equations, a system of d linear equations in d unknowns.
  - Time complexity (naive) of Gaussian elimination algorithm: $O(d^3)$.

Classical statistics view of OLS (1)

- *Normal linear regression model*
- Model training examples $(X_1, Y_1), \ldots, (X_n, Y_n)$ as iid random variables taking values in $\mathbb{R}^d \times \mathbb{R}$ where

$$Y \quad X = x \qquad \mathrm{N}(x^\mathsf{T} w, \sigma^2)$$

-
- problem of finding the maximum likelihoo

Classical statistics view of OLS (2)

- Suppose your data really does come from a distribution in this statistical model, say, with parameters w and σ^2.
  - Then the function with smallest MSE is the linear function $f^\star(x) = x^\intercal w$, and its MSE is $\mathrm{mse}(f^\star) = \sigma^2$.
  - So estimating w is a sensible idea! (Plug-in principle...)

Statistical learning view of OLS (1)

- IID model: $(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y) \sim_{\text{iid}} P$ are iid random variables taking values in $\mathbb{R}^d \times \mathbb{R}$
- (X, Y) is the (unseen) 'test' example
- Goal: find a (linear) function $w \in \mathbb{R}^d$ with small MSE
- $w \in \mathbb{R}^d$,

since it is an expectation (e.g., integral) with r
unknown distribution P

Statistical learning view of OLS (2)

- However, we have an iid sample $S := ((X_1, Y_1), \ldots, (X_n, Y_n))$.
- We swap out P in the definition of $\mathrm{mse}(f)$, and replace it with the *empirical distribution* on S

$$\frac{1}{n}$$

on the i-th training example.

- Resulting objective function is

$$\mathbb{E}[(\tilde{X}^\mathsf{T} w - \tilde{Y})^2] = \frac{1}{n} \sum_{i=1} (X_i^\mathsf{T} w - Y_i)$$

where $(\tilde{X}, \tilde{Y}) \sim P_n$.

Statistical learning view of OLS (3)

- In some circles:
  - *(True/population) risk* of w: $\mathcal{R}(w) := \mathbb{E}[(X^\intercal w - Y)^2]$
  - *Empirical risk* of w: $\widehat{\mathcal{R}}(w) := \frac{1}{n}\sum_{i=1}^{n}(X_i^\intercal w - Y_i)^2$
- This is another instance of the plug-in principle!

Statistical learning view of OLS (4)

- This is not specific to linear regression; also works for other types of functions, and also other types of prediction problems, including classification.
- For classification:

- Procedure that minimizes empirical risk:
  Empirical risk minimization (*ERM*)

Upgrading linear regression (1)

- Make linear regression more powerful by being creative about features
  - We are forced to do this if x is not already provided as a vector of numbers
- ion φ

Upgrading linear regression (2)

- Examples:
  - Affine feature expansion, e.g., $\varphi(x) = (1, x)$, to accommodate intercept
  - Standardization, e.g., $\varphi(x) = (x - \mu)/\sigma$ where (μ, σ^2) are (estimates of) the mean and variance of the feature value
  - Polynomial expansion, e.g., $\varphi(x) = (1, x_1, \ldots, x_d, x_1^2, \ldots, x$
  - Headless neural network $\varphi(x) =$ $N : \mathbb{R}^d \to \mathbb{R}^k$ is a map computed by a in neural network
    - (Later, we'll talk about how to "learn" N.)

Example: Taking advantage of linearity

- Example: y is health outcome, x is body temperature
  - Physician suggests relevant feature is (square) deviation from normal body temperature $(x - 98.6)^2$
  - What if you didn't know the magic constant 98.6? (Apparently it is wrong in the US anyway)

Example: Binning features

- Dartmouth data example, where we considered intervals for the HS GPA variable:

$$(0.00, 0.25], \quad (0.25, 0.50], \quad (0.50, 0.75], \quad \ldots$$

- ear
-
- $\varphi(x)^\top w = w_j$ if x is in the j-th i

Effect of feature expansion on expected MSE

$$\mathbb{E}[\text{mse}(\hat{f})] = \underbrace{\mathbb{E}\,\text{var}(Y \mid X)}_{\text{u}} + \mathbb{E}\Big[\text{var}(\hat{f}(X) \mid X)\Big] + \underbrace{\mathbb{E}\Big[(\mathbb{E}[\hat{f}(X) \mid X] - \mathbb{E}[Y \mid X])^2\Big]}_{\text{n error of } \hat{f}}$$

-

(approximation error)

- But maybe at the cost of increasing the secon (variability)

Performance of OLS (1)

- Study in context of IID model
- $(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)$ are iid, and assume $\mathbb{E}[XX^\intercal]$ is invertible (WLOG).
- Let w^* denote the minimizer of $\text{mse}(w)$ over all $w \in \mathbb{R}^d$.

- How much larger is $\text{mse}(\hat{w})$ comp

Performance of OLS (2)

- **Theorem**: In the IID model, the OLS solution $\hat{w}$ satisfies

$$n\left(\mathbb{E}[\mathrm{mse}(\hat{w})] - \mathrm{mse}(w^*)\right) \to \mathrm{tr}(\mathrm{cov}(\varepsilon W))$$

as n , where $W = \mathbb{E}[XX^\intercal]^{-1/2}X$ and $\varepsilon = Y \quad X^\intercal w^*$.

-

$$Y \mid X \quad x \sim \quad x \; w \; , \sigma$$

$$n\left(\mathbb{E}[\mathrm{mse}(\hat{w})] - \mathrm{ns}\right)$$

which is more typically written as

$$\mathbb{E}[\mathrm{mse}(\hat{w})] \to \left(1 + \frac{d}{n}\right) \mathrm{mse}(w^*).$$

Linear algebraic view of OLS (1)

- Write $A = \begin{bmatrix} \uparrow & & \uparrow \\ a_1 & \cdots & a_d \\ \downarrow & & \downarrow \end{bmatrix}$
  - $a_j \in \mathbb{R}^n$ is j-th column of A
    - Span of $a_1, \ldots, a_d$ is $\text{range}(A)$, a subspace of $\mathbb{R}^n$
- $\in$

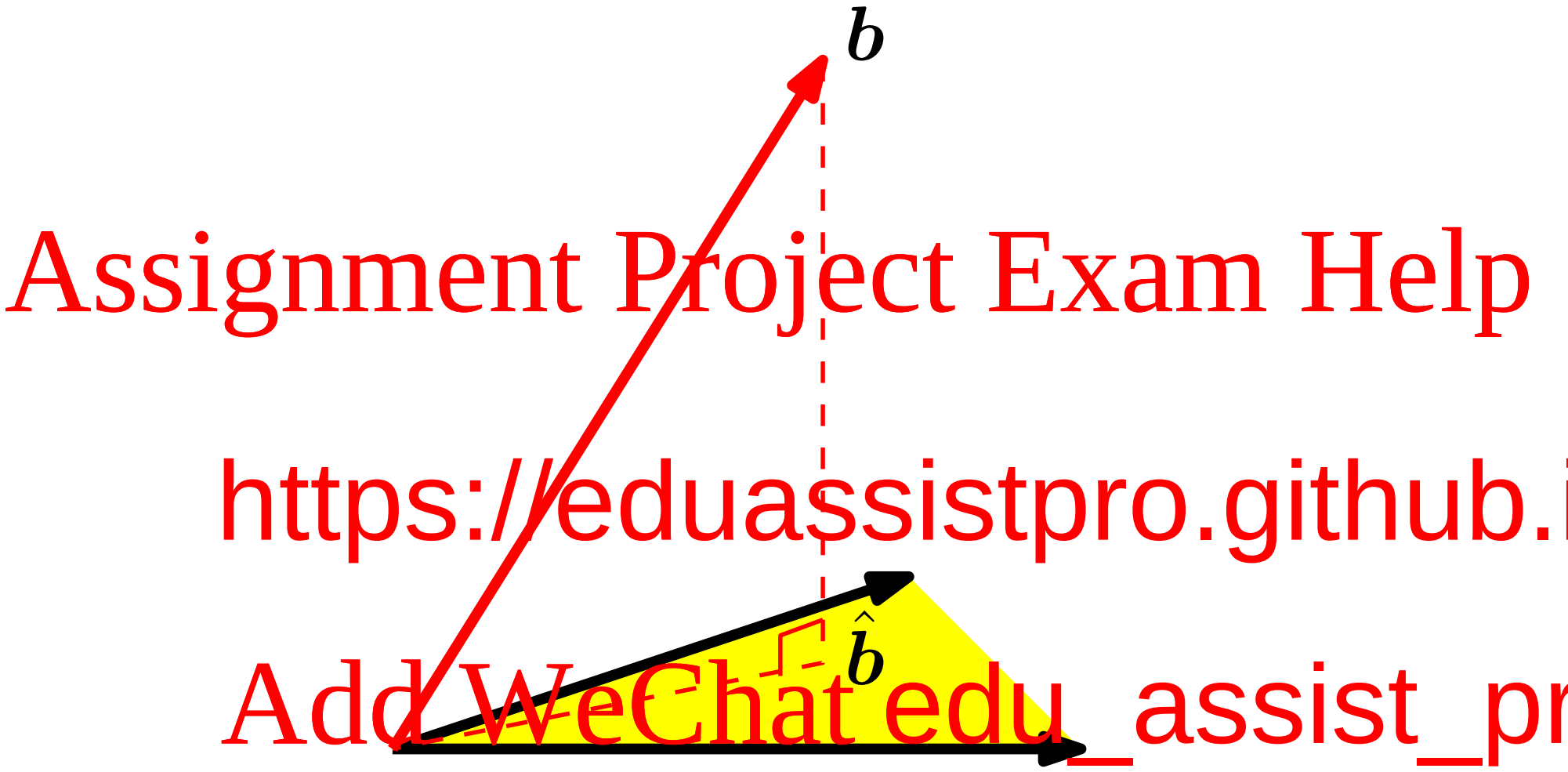

Figure 8: Orthogonal projection of b onto $\mathrm{range}(A)$

Linear algebraic view of OLS (2)

- Solution $\hat{b}$ is *orthogonal projection* of b onto $\mathrm{range}(A)$
  - $\hat{b}$ is unique
  - *Residual* $b - \hat{b}$ is orthogonal to $\hat{b}$
  - To get w from $\hat{b}$, solve $Aw = \hat{b}$ for w.
  - If $\mathrm{rank}(A) < d$ (always the case if $n < d$), then infinitely-many
- $n \geq d$,

Over-fitting (1)

- In the IID model, *over-fitting* is the phenomenon where the true risk is much worse than the empirical risk.

Over-fitting (2)

- Example:
  - $\varphi(x) = (1, x, x^2, \ldots, x^k)$, degree-$k$ polynomial expansion
  - Dimension is $d = k + 1$
  - Any function of $\leq k + 1$ points can be interpolated by polynomial of degree $\leq k$

ero

).

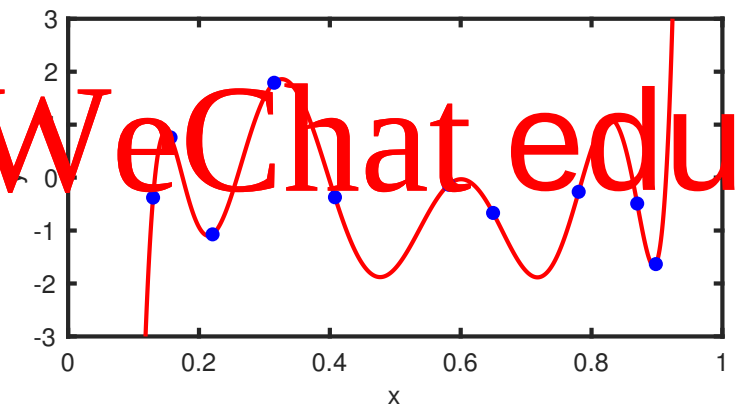

Figure 9: Polynomial interpolation

Beyond empirical risk minimization

- Recall plug-in principle
  - Want to minimize risk with respect to (unavailable) P; use P_n instead
- What if we can't regard data as iid from P?

$$\underline{1} \qquad \underline{1}$$

e from

- How to implement plug-in principle?