# Machine learning lecture slides

COMS 4771 Fall 2020

# Regression I: Linear regression

# Outline

- Statistical model for regression
- College GPA example
- Ordinary least squares for linear regression
- The expected mean squared error
- Different views of ordinary least squares
- Features and linearity
- Over-fitting
- Beyond empirical risk minimization

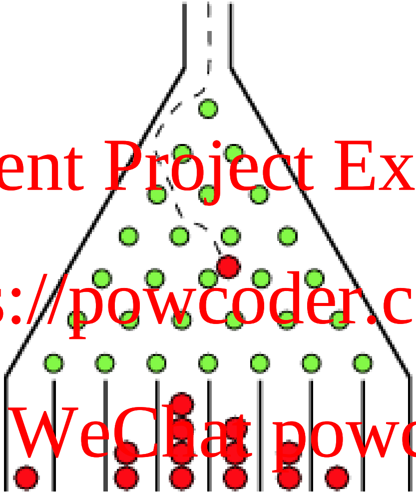



Figure 1: Galton board

# Real-valued predictions

- Example: Galton board
- Physical model: hard
- Statistical model: final position of ball is random
  - *Normal (Gaussian) distribution* with mean $\mu$ and variance $\sigma^2$
  - Written $\mathrm{N}(\mu, \sigma^2)$
  - Probability density function is

$$p_{\mu,\sigma^2}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \quad y \in \mathbb{R}.$$

- Goal: predict final position accurately, measure *squared loss* (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*:



$$\begin{aligned}
\mathbb{E}[(\hat{y} - Y)^2] &= \mathbb{E}[(\hat{y} - \mu + \mu - Y)^2] \\
&= (\hat{y} - \mu)^2 + 2(\hat{y} - \mu)\mathbb{E}[(\mu - Y)] + \mathbb{E}[(\mu - Y)^2] \\
&= (\hat{y} - \mu)^2 + \sigma^2.
\end{aligned}$$



  - This is true for any random variable $Y$; don't need normality 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 $f$:

$$\mathrm{mse}(f) := \mathbb{E}[(f(X) - Y)^2],$$

the expected squared loss of $f$ on random example

# Optimal prediction function for regression

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

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

  - Also called the *regression function* or *conditional mean function*
  - Prediction function with smallest MSE
  - Depends on conditional distribution of $Y$ given $X$

- 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:
  $(X_1, Y_1), \dots, (X_n, Y_n), (X_1', Y_1'), \dots, (X_m', Y_m'), (X, Y)$ are iid
  - Training examples (that you have):
    $S := ((X_1, Y_1), \dots, (X_n, Y_n))$
  - Test examples (that you have): $T := ((X_1', Y_1'), \dots, (X_m', Y_m'))$
  - Test example (that you don't have) used to define MSE: $(X, Y)$
- Predictor $\hat{f}$ is based only on training examples
- Hence, **test examples are independent of** $\hat{f}$ (very important!)
- We would like to estimate $\mathrm{mse}(\hat{f})$

# Test MSE (2)

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

- 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)
- Absent any other features, best constant prediction of a uniformly random Dartmouth student's College GPA is $y := 2.46$.



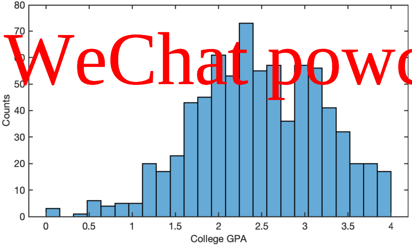


Figure 2: Histogram of College GPA

# 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
  - $y_i$ is College GPA of $i$-th student

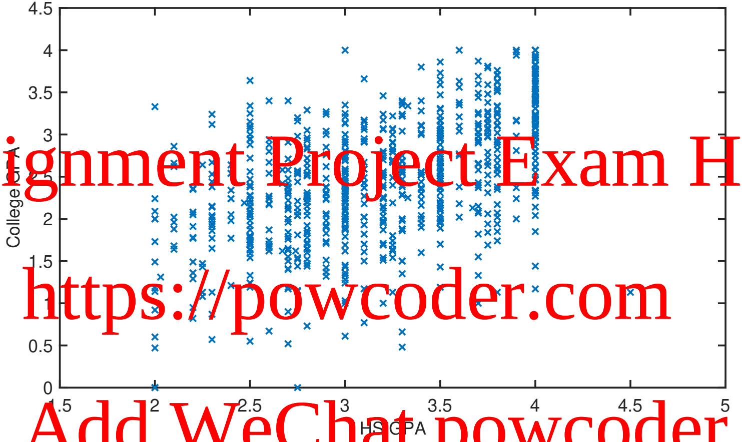



Figure 3: Plot of College GPA vs HS GPA

## Predicting College GPA from HS GPA (2)

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

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

  - For each such interval $I$, record the mean $\hat{\mu}_I$ of the College GPAs of students whose HS GPA falls in $I$.

$$\hat{f}(x) := \begin{cases} \hat{\mu}_{(0.00,0.25]} & \text{if } x \in (0.00, 0.25] \\ \hat{\mu}_{(0.25,0.50]} & \text{if } x \in (0.25, 0.50] \\ \hat{\mu}_{(0.50,0.75]} & \text{if } x \in (0.50, 0.75] \\ & \vdots \end{cases}$$

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

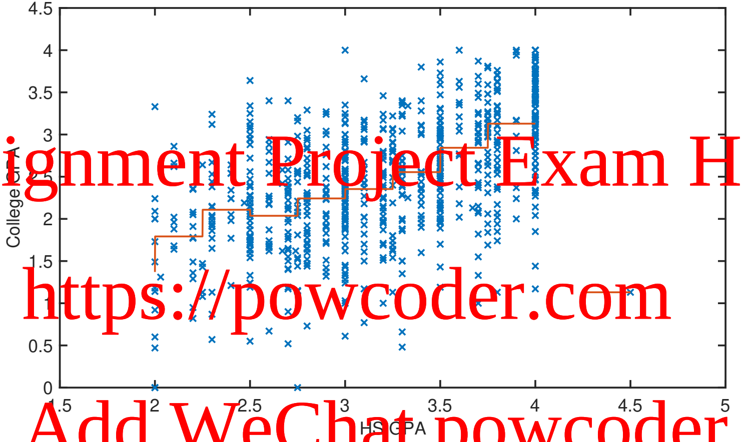



Figure 4: Plot of mean College GPA vs binned HS GPA

## Predicting College GPA from HS GPA (3)

- Define
$$\mathrm{mse}(f, S) := \frac{1}{|S|} \sum_{(x,y)\in S} (f(x) - y)^2,$$
the mean squared error of predictions made by $f$ on examples in $S$.
  - "mean" is with respect to the uniform distribution on examples in $S$.

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

$$\sqrt{\mathrm{mse}(\hat{f}, S)} = 0.613 < 0.746 \text{ (the standard deviation of the } y_i\text{'s)}$$

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

- 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 4.5



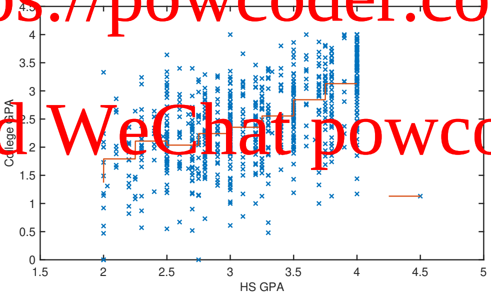


Figure 5: Plot of mean College GPA vs binned HS GPA

## Least squares linear regression (1)

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

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

for $m, \theta \in \mathbb{R}$.

  - Technically, $x \mapsto mx + \theta$ is linear iff $\theta = 0$. If $\theta \neq 0$, the function is not linear but *affine*.
  - Semantics: Positive $m$ means higher HS GPA gets a higher 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

$$\frac{1}{n} \sum_{i=1}^{n} (mx_i + \theta - y_i)^2.$$

- Also called *ordinary least squares* (*OLS*)

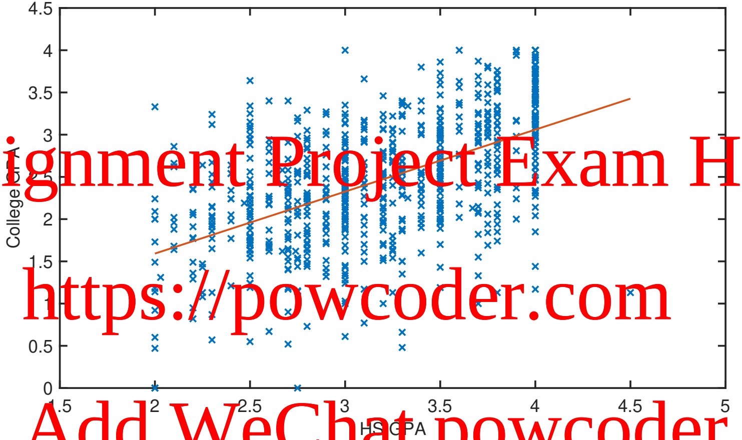



Figure 6: Plot of least squares linear regression line

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

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

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



  - System of two linear equations with two unknowns $(m, \theta)$.
  - Define

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{x^2} := \frac{1}{n}\sum_{i=1}^{n} x_i^2,$$

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

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:

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



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

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

- 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 the LHS of system of 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 equations is an optimal solution to the least squares linear regression problem.

- 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.
- IID model: $(X_1, Y_1), \ldots, (X_n, Y_n), \quad (X, Y)$ are iid
- Say training examples $(X_1, Y_1), \ldots, (X_n, Y_n)$ are used to determine $\hat{f}$.
- What is $\mathbb{E}[\text{mse}(\hat{f})]$?

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

# 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:
  - probability distribution of training data,
  - type of function being fit (e.g., piecewise constant, linear),
  - method of fitting (e.g., OLS),
  - etc.
- Third term quantifies how well a function produced by the fitting procedure can approximate the regression function, even after removing the "variability" of $\hat{f}$.

# 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).
  - Linear regression: a function $f \colon \mathbb{R}^2 \to \mathbb{R}$ of the form

$$f(x) = m_1 x_1 + m_2 x_2 + \theta$$

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

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

$$f(x) = x^{\mathsf{T}} w$$



for some $w \in \mathbb{R}^d$.
  - $w$ is called the *weight vector* or *coefficient vector*.
  - What about inhomogeneous linear functions?
    - Just always include a "feature" that always has value $1$. Then the corresponding weight acts like $\theta$ from before.

# Multivariate ordinary least squares (1)

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

$$\widehat{\mathcal{R}}(w) := \frac{1}{n} \sum_{i=1}^{n} (x_i^{\mathsf{T}} w - y_i)^2.$$

  - Notation warning: $x_i \in \mathbb{R}^d$

# Multivariate ordinary least squares (2)

- In matrix notation:

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



where

$$A := \frac{1}{\sqrt{n}} \begin{bmatrix} \longleftarrow & x_1^\mathsf{T} & \longrightarrow \\ & \vdots & \\ \longleftarrow & x_n^\mathsf{T} & \longrightarrow \end{bmatrix} \in \mathbb{R}^{n \times d}, \quad b := \frac{1}{\sqrt{n}} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n.$$



- If we put vector $v \in \mathbb{R}^d$ in the context of matrix multiplication, it is treated as a column vector by default!
  - If we want a row vector, we write $v^\mathsf{T}$.
- Therefore

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

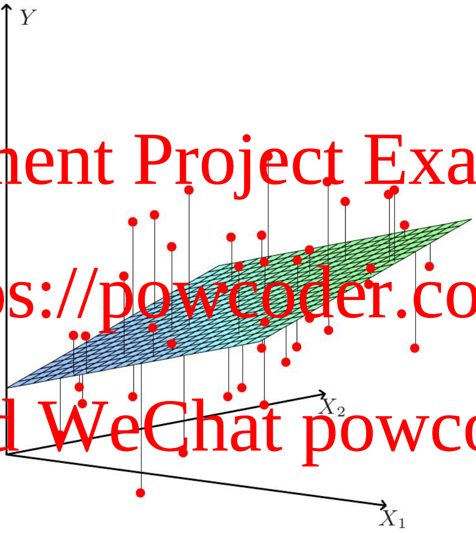



Figure 7: Geometric picture of least squares linear regression

# 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*:

$$\nabla_w \widehat{\mathcal{R}}(w) = \begin{bmatrix} \frac{\partial \widehat{\mathcal{R}}(w)}{\partial w_1} \\ \vdots \\ \frac{\partial \widehat{\mathcal{R}}(w)}{\partial w_d} \end{bmatrix} = 2A^{\mathsf{T}}(Aw - b) = 0,$$

which is equivalent to

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

## Multivariate normal equations (2)

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



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

- If $A^{\mathsf{T}}A$ is singular, then there are infinitely many solutions!





- **Theorem**: Every solution to the normal equations is an optimal solution to the least squares linear regression problem.

## 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: $O(d^3)$.
  - Actually, also need to count time to form the system of equations, which is $O(nd^2)$.

- *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



$$Y_i \mid X_i = x_i \sim \mathrm{N}(x_i^\intercal w, \sigma^2)$$



  - $w \in \mathbb{R}^d$ and $\sigma^2 > 0$ are the parameters of the model.
- The least squares linear regression problem is the same as the problem of finding the maximum likelihood value for $w$.

## Classical statistics view of OLS (2)

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

- IID model: $(X_1, Y_1), \dots, (X_n, Y_n), (X, Y) \sim_{\text{iid}} P$ 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

$$\text{mse}(w) = \mathbb{E}[(X^\top w - Y)^2].$$

- We cannot directly minimize $\text{mse}(w)$ as a function of $w \in \mathbb{R}^d$, since it is an expectation (e.g., integral) with respect to the unknown distribution $P$

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



$$P_n(x, y) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{(x,y)=(x_i,y_i)\}}.$$



  - This is the distribution that puts probability mass $1/n$ on the $i$-th training example.
- Resulting objective function is



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

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

- In some circles:
  - *(True/population) risk* of $w$: $\mathcal{R}(w) := \mathbb{E}[(X^\top w - Y)^2]$
  - *Empirical risk* of $w$: $\widehat{\mathcal{R}}(w) := \frac{1}{n}\sum_{i=1}^{n}(X_i^\top w - Y_i)^2$
- This is another instance of the plug-in principle!
  - We want to minimize $\mathrm{mse}(w)$ but we don't know $P$, so we replace it with our estimate $P_n$.

- 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:
  - *(True/population) risk* of $f$: $\mathcal{R}(f) := \mathbb{E}[\mathbf{1}_{\{f(X)\neq Y\}}]$
  - *Empirical risk* of $f$: $\widehat{\mathcal{R}}(f) := \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{f(X_i)\neq Y_i\}}$
  - All that changed is the *loss function* (squared loss versus *zero/one loss*)
- 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
- Instead of using $x$ directly, use $\varphi(x)$ for some transformation $\varphi$ (possibly vector-valued)

# 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 $(\mu, \sigma^2)$ are (estimates of) the mean and variance of the feature value
  - Non-linear scalar transformations, e.g., $\varphi(x) = \ln(1 + x)$
  - Logical formula, e.g., $\varphi(x) = (x_1 \wedge x_5 \wedge \neg x_{10}) \vee (\neg x_2 \wedge x_7)$
  - Trigonometric expansion, e.g., $\varphi(x) = (1, \sin(x), \cos(x), \sin(2x), \cos(2x), \dots)$
  - Polynomial expansion, e.g., $\varphi(x) = (1, x_1, \dots, x_d, x_1^2, \dots, x_d^2, x_1 x_2, \dots, x_{d-1} x_d)$
  - *Headless neural network* $\varphi(x) = N(x) \in \mathbb{R}^k$, where $N \colon \mathbb{R}^d \to \mathbb{R}^k$ is a map computed by a intermediate layer of a 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)
  - Use $\varphi(x) = (1, x, x^2)$
  - Can learn coefficients $w \in \mathbb{R}^3$ such that $\varphi(x)^\top w = (x - 98.6)^2$, or any other quadratic polynomial in $x$ (which could be better!)

# Example: Binning features

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

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

- Use $\varphi(x) = (\mathbf{1}_{\{x \in (0.00,0.25]\}}, \mathbf{1}_{\{x \in (0.25,0.50]\}}, \dots)$ with a linear function
- What is $\varphi(x)^\intercal w$?
  - $\varphi(x)^\intercal w = w_j$ if $x$ is in the $j$-th interval.

# Effect of feature expansion on expected MSE



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



- Feature expansion can help reduce the third term (approximation error)
- But maybe at the cost of increasing the second term (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 $\mathbb{E}[XX^\mathsf{T}]$ is invertible (WLOG).
- Let $w^*$ denote the minimizer of $\mathrm{mse}(w)$ over all $w \in \mathbb{R}^d$.
  - Inductive bias assumption: $\mathrm{mse}(w^*)$ is small, i.e., there is a linear function with low MSE.
  - This is a fairly "weak" modeling assumption, especially compared to the normal regression model.
- How much larger is $\mathrm{mse}(\hat{w})$ compared to $\mathrm{mse}(w^*)$?

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



- **Corollary**: If, in addition, $(X, Y)$ follows the normal linear regression model $Y \mid X = x \sim \mathrm{N}(x^\intercal w^*, \sigma^2)$, then

$$n\left(\mathbb{E}[\mathrm{mse}(\hat{w})] - \mathrm{mse}(w^*)\right) \to \sigma^2 d$$



which is more typically written as

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

- 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 $a_1, \dots, a_d$ is $\mathrm{range}(A)$, a subspace of $\mathbb{R}^n$
- Minimizing $\widehat{\mathcal{R}}(w) = \|Aw - b\|_2^2$ over $w \in \mathbb{R}^d$ is same as finding vector $\hat{b}$ in $\mathrm{range}(A)$ closest to $b$

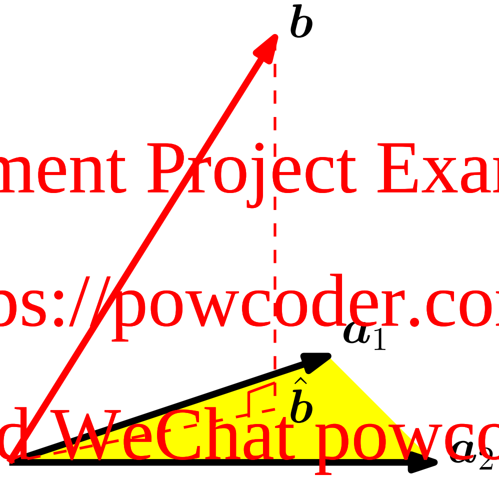


Figure 8: Orthogonal projection of $b$ onto $\mathrm{range}(A)$

# 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 ways to write $\hat{b}$ as linear combination of $a_1, \ldots, a_d$.
- Upshot: Uniqueness of least squares solution requires $n \geq d$, and $n < d$ guarantees non-uniqueness!

## Over-fitting (1)

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

- 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 $\leq k$
  - So if $n \leq k + 1 = d$, least squares solution $\hat{w}$ will have zero empirical risk, regardless of its true risk (assuming no two training examples with distinct $x_i$'s have different $y_i$'s).



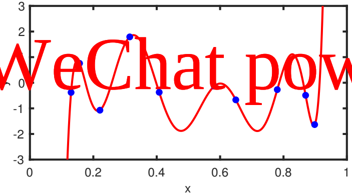


Figure 9: Polynomial interpolation

## 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 $P$?
  - Example: Suppose we know $P = \frac{1}{2}M + \frac{1}{2}F$ (*mixture distribution*)
  - We get size $n_1$ iid sample from $M$, and size $n_2$ iid sample from $F$, $n_2 \ll n_1$
  - How to implement plug-in principle?