1/26/2023: Overfitting, Regularization

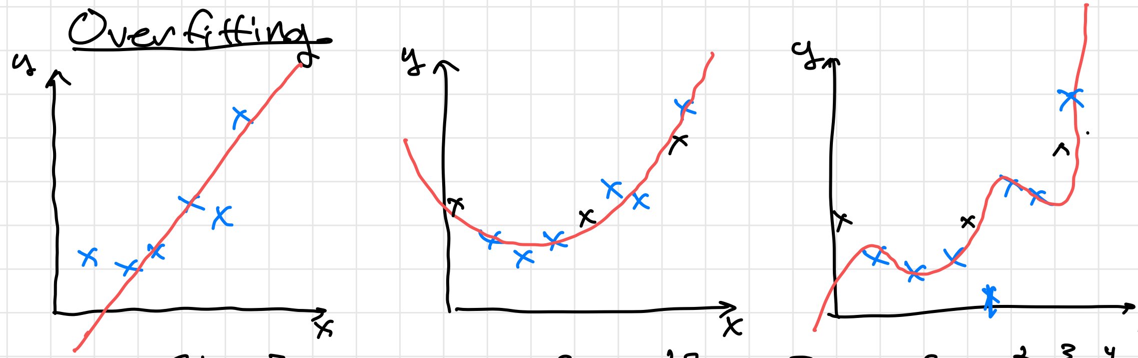
Review of linear supervised learning methods so far

supervised: learn from dataset of (x, y) pairs

input (x); correct answer "supervision" (y)

<table>
<tr><th></th><th>Linear Regression</th><th>Logistic Regression</th><th>Softmax Regression</th></tr>
<tr><td>Task</td><td>Regression
$y \in \mathbb{R}$</td><td>Binary classification
$y \in \{-1, 1\}$</td><td>multiclass classification
$y \in \{1, 2, 3, \ldots, C\}$</td></tr>
<tr><td>Parameters
(what to learn)</td><td>$w \in \mathbb{R}^{\boxed{d}}$
dimension of x</td><td>$w \in \mathbb{R}^d$</td><td>$w^{(1)}, \ldots, w^{(C)} \in \mathbb{R}^d$
total $C \cdot d$ params</td></tr>
<tr><td>Probabilistic story</td><td>$y \sim \text{Normal}(w^T x, \sigma^2)$
mean</td><td>$p(y=1 \mid x) = \sigma(w^T x)$
</td><td>$p(y=j) = \dfrac{\exp(w^{(j)T} x)}{\boxed{\sum_{k=1}^{C} \exp(w^{(k)T} x)}}$
Normalizes to probability distribution</td></tr>
<tr><td>How to get loss function
measures how bad any choice of parameters is</td><td colspan="3">Maximum Likelihood Estimation (MLE)
maximize probability of data $= \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; w)$ with respect to w
$\iff$ minimize negative log likelihood $= -\sum_{i=1}^{n} \log p(y^{(i)} \mid x^{(i)}; w)$</td></tr>
<tr><td>How to minimize loss</td><td>Gradient descent OR Normal Equations</td><td colspan="2">Gradient descent (1st-order)
OR
Newton-Raphson Method (2nd order) (also uses 2nd derivs)</td></tr>
</table>

Overfitting



Features = [1, x] | Features = [1, x, x^2] | Features = [1, x, x^2, x^3, x^4, x^5]

Too simple
"underfitting"

☆ Best fit

NOT THE SAME

Sensitive to fluctuations in training
"overfitting"

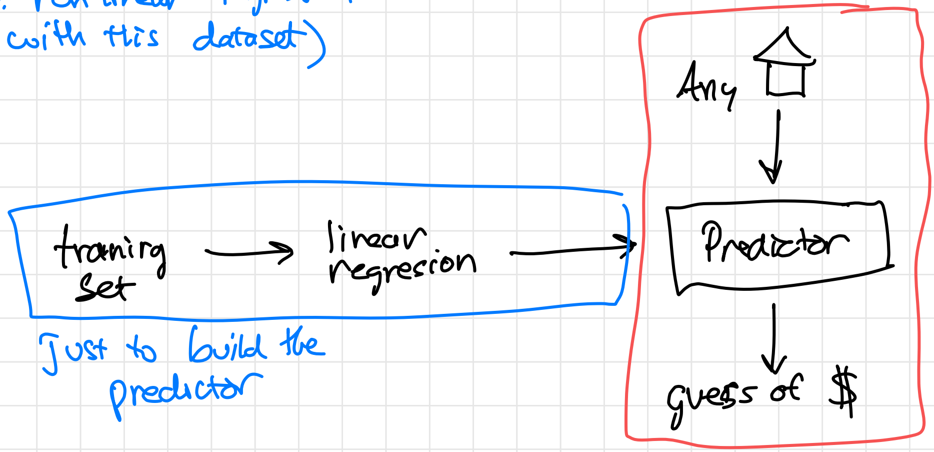
Lowest training loss

Dataset Splits: ALWAYS have 3 datasets with disjoint examples

Training set
Use it to choose parameters
(e.g. run linear regression with this dataset)

Development Set
("Validation")

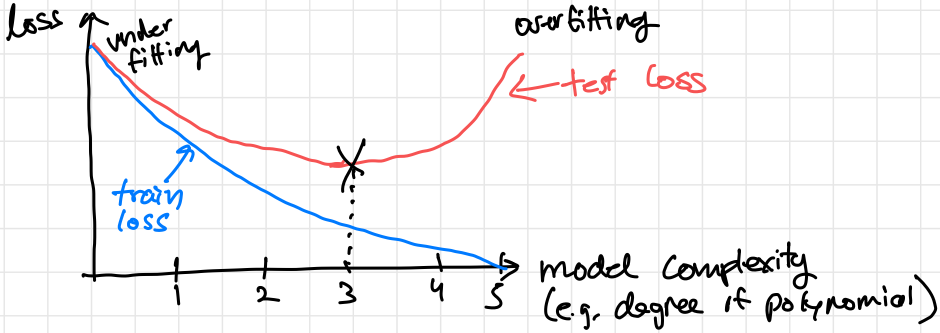
Test set
Evaluate how well our model does



Just to build the predictor

What matters: accuracy of predictions on unseen examples

Test set simulates model behavior on unseen examples



What is development set for?

A: To choose hyperparameters

Any setting of learning algorithm
- Learning rate
- Which features?
- When to stop training

Rule: Choose hyperparameters to minimize development set loss,

(In contrast, parameter is chosen by the learning algorithm)

only evaluate on test set at very end.

degree	dev loss	final evaluation test loss
1	100	100
2	51	50
3	50	50
$\boxed{4}$	$\boxed{49}$	50
5	75	75

1. Train 5 models
2. Evaluate each on dev set
3. Pick the model that's best on dev
4. Evaluate that model on test set

Announcements
- HW 1 released Feb 7
- HW 0 grades out

Test questions

Alice's test score

Bob test score

"Correlation does not imply Causation"

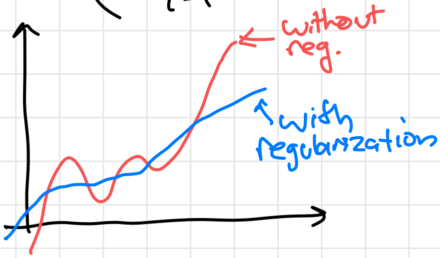
Regularization: A way to reduce overfitting by preferring "simpler" functions

L_2 Regularization: Encourage square of $\boxed{L_2 \text{ norm}}$ of parameters to be small

$$\sum_{j=1}^{d} w_j^2 = \|w\|^2$$

e.g. linear regression

$$L(w) = \left(\frac{1}{n} \sum_{i=1}^{n} (w^T x^{(i)} - y^{(i)})^2 \right) + \lambda \|w\|^2$$



some constant hyperparameter

$\lambda = 0 \Rightarrow$ no regularization

λ large $\Rightarrow$ strong regularization

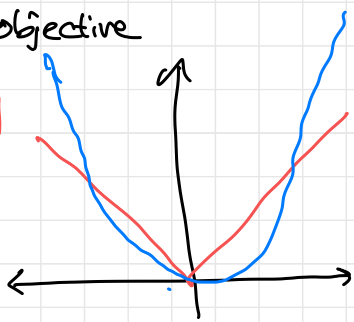
How does this change the gradient

gradient of $\lambda \|w\|^2 = 2\lambda w$

during GD, you subtract $\eta \cdot 2\lambda w$ "weight decay"

learning rate

L_1 regularization: add $\lambda \underbrace{\|w\|_1}_{:= \sum_{j=1}^{d} |w_j|}$ to objective



Gradient for L_1:

$$\frac{d}{dw_j} \lambda \|w\|_1 = \lambda \, \mathrm{sgn}(w_j)$$

$$\mathrm{sgn}(z) = \begin{cases} 1 & \text{if } z > 0 \\ -1 & \text{if } z < 0 \\ 0 & \text{if } z = 0 \end{cases}$$

So full gradient is $\lambda \begin{bmatrix} \mathrm{sgn}(w_1) \\ \vdots \\ \mathrm{sgn}(w_d) \end{bmatrix}$ vs $2\lambda w$ from L_2

constant sized step towards 0

if w close to 0, you take very small step

L_1 regularization has a sparsifying effect (leads to $\underbrace{\text{sparse}}_{}$ w)

means many entries $= 0$

L_2 avoids very big entries of w