Linear Regression

Program:

$$\underset{w}{\operatorname{argmin}} \frac{1}{2} \| \underbrace{Y}_{\mathbb{R}^N} - \underbrace{X^T}_{\mathbb{R}^{N \times d+1}} \underbrace{w}_{\mathbb{R}^{d+1}} \|_2^2$$

Solution:

$$w^* = (X X^T)^{-1} X Y$$

With Regularization:

Prog: $\underset{w}{\operatorname{argmin}} \frac{1}{2} \| Y - X^T w \|_2^2 + \frac{C}{2} \| w \|_2^2$

Solution: $w^* = (X X^T + CI)^{-1} X Y$

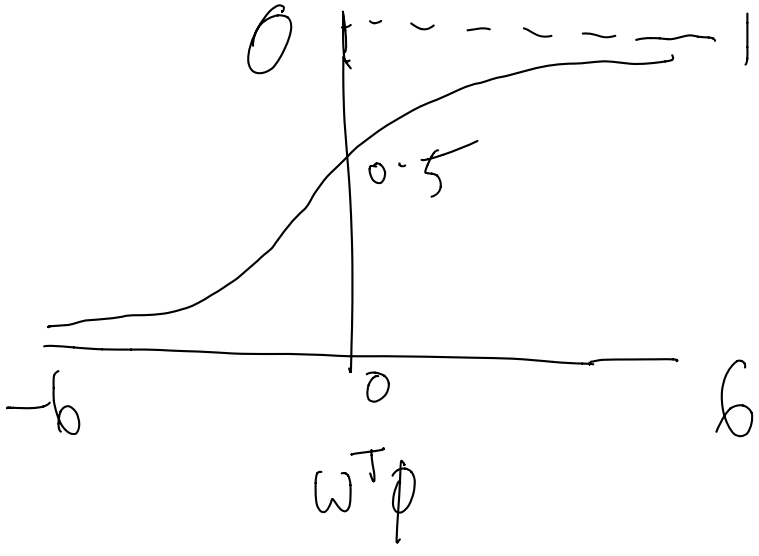
L1: Lasso → Shrinks features to 0 (feature selection)

L2: Ridge → Shrinks features but doesn't remove them.

Logistic Regression

- Used in Classification
- Does NOT penalize when samples are easy to classify
- Sigmoid Function

$$\left(1 + \exp(-w^T \phi)\right)^{-1}$$



- Squishes output to ~~be~~ between 0 and 1

- Model:

$$y^{(i)} \in \{-1, 1\}$$

$$P(y^{(i)} = 1 \mid x^{(i)}; w) = \frac{1}{1 + \exp(-w^T \phi)}$$

$$P(y^{(i)}=1) = \frac{1}{1+\exp(-w^T\phi)}$$

$$P(y^{(i)}=-1) = 1 - P(y^{(i)}=1) = \frac{1}{1+\exp(w^T\phi)}$$

$$\boxed{P(y^{(i)}|x^{(i)}) = \frac{1}{1+\exp(-y^{(i)}w^T\phi)}}$$

- Maximum Likelihood.

$$\arg\max_w \prod_{(x^{(i)},y^{(i)})\in D} P(y^{(i)}|x^{(i)})$$

$$= \arg\min_w -\prod P(y^{(i)}|x^{(i)})$$

$$= \arg\min_w \sum_{|D|} -\log\left(P(y^{(i)}|x^{(i)})\right)$$

$$\boxed{= \arg\min \sum \log\left(1+\exp(-y^{(i)}w^T\phi)\right)}$$

$$\nabla_w \sum_{|D|} \log(1 + \exp(\ldots$$

- Optimize Via Gradient Descent

$$\nabla_w = \sum_{|D|} \frac{-y^{(i)} \exp(-y^{(i)} w^T \phi)}{1 + \exp(-y^{(i)} w^T \phi)} \phi$$

Initialize $t=0$, w_t and α

Iterate:

$$g_t = \nabla_w f(w_t)$$

$$w_{t+1} = w_t - \alpha g_t$$

$$t \leftarrow t+1$$

Optimization

When Can we find optimum?

- Linear, least squares, Convex etc
- Generally can be hard
- Jensen's Inequality:

Convex. $f((1-\lambda)w_1 + \lambda w_2) \leq (1-\lambda)f(w_1) + \lambda f(w_2$

- $\nabla^2 f(w) \geq 0 \ \forall \ w$
- Batch Vs Stochastic
  - Batch looks at whole data set
  - Stochastic looks at a part of the data set.

$\min_w f_0(w)$

s.t. $f_i(w) \leq 0 \quad \forall i \in \{1, \dots, c_1\}$

$h_i(w) = 0 \quad \forall i \in \{1, \dots, c_2\}$

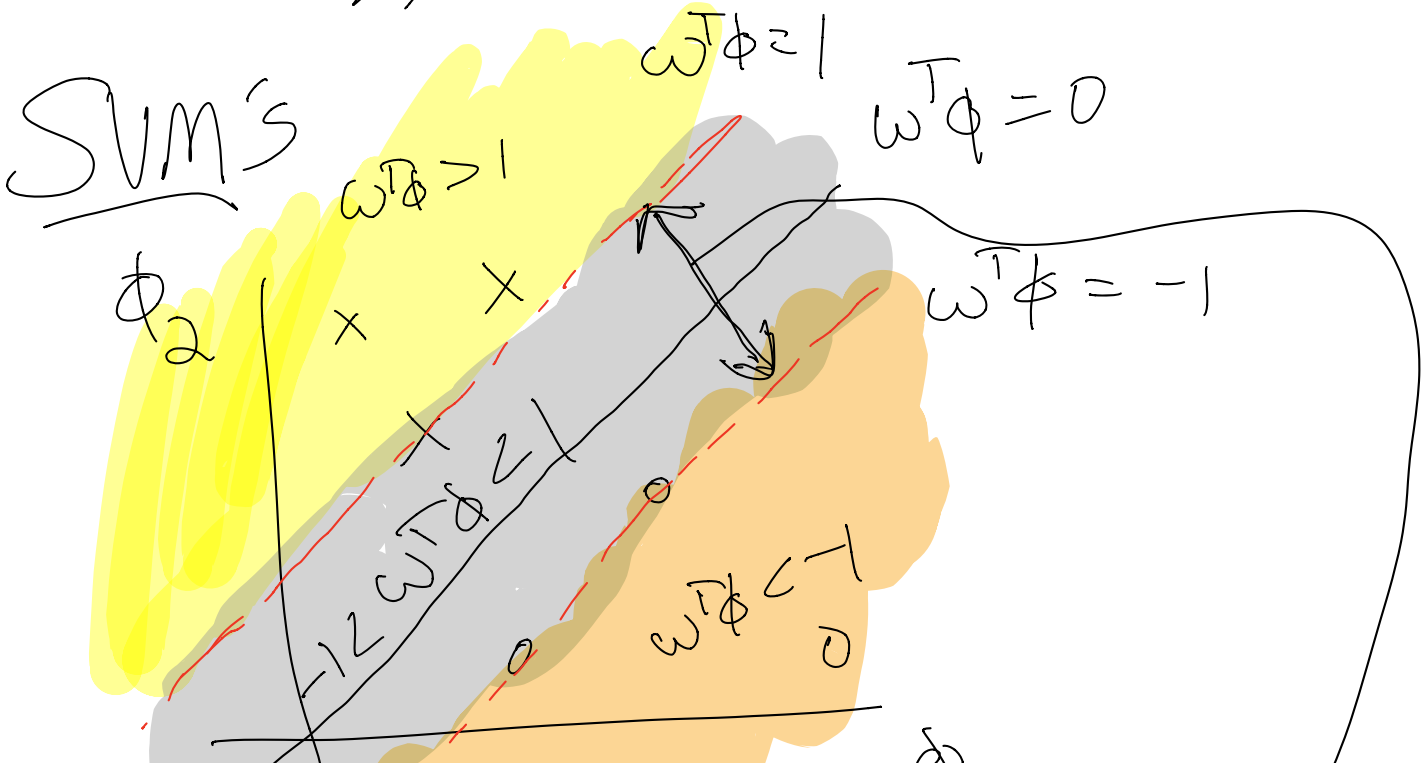
$$L(w, \lambda, \nu) = f_0(w) + \sum_{i=1}^{c_1} \lambda_i f_i(w) + \sum_{i=1}^{c_2} \nu_i h_i(w).$$

$\lambda_i \geq 0$

$$f(w) \geq L \geq \min_w L = g(\lambda, \nu)$$

$$\max_{\lambda, \nu} g(\lambda, \nu) \text{ ST } \lambda_i \geq 0 \ \forall i$$

SVM's

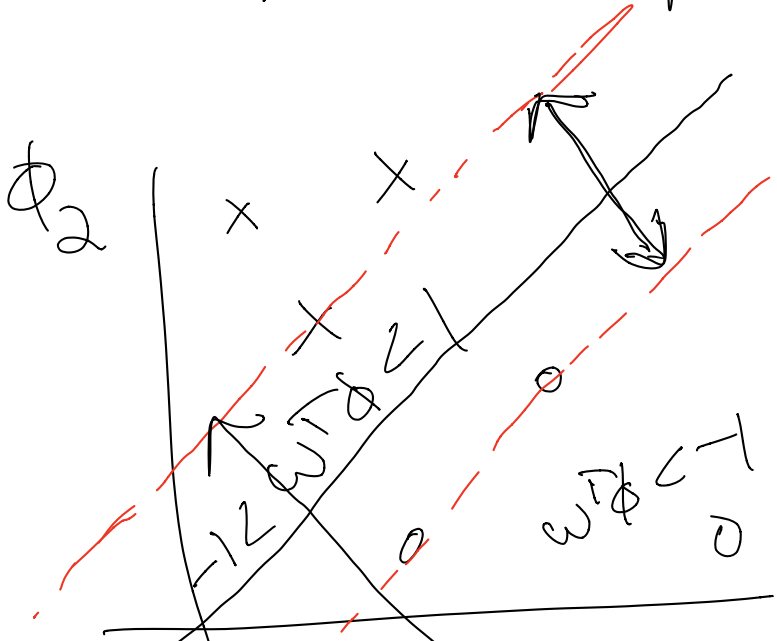


goal of SVM is to max this margin (seperate data the most).

$$\text{margin} = \frac{2}{\|w\|}$$

$$\max_w \frac{2}{\|w\|} = \boxed{\min_w \frac{1}{2}\|w\|_2^2 \quad ST \quad y_i w^T \phi \geq 1}$$

- Slack Variable (ξ)
  - for linearly unseperable data.



$$\min_{w,\xi \geq 0} \frac{C}{2}\|w\|_2^2 + \sum_{i \in D} \xi^{(i)}$$

$$\text{s.t.} \quad y^{(i)} w^T \phi \geq 1 - \xi^{(i)}$$

$$\xi \geq 1 - y^i w^T \phi, \quad \xi \geq 0$$

$$\min_w \frac{C}{2}\|w\|_2^2 + \sum \max(0, 1 - y^{(i)} w^T \phi)$$

takes care of $\xi \geq 0$

$$L(\cdot) = \frac{C}{2}\|w\|_2^2 + \sum_i \xi^i + \sum_i \alpha^i (1 - y^i w^T \phi - \xi^i)$$

$$= \frac{C}{2}\|w\|_2^2 - w^T \sum_i \alpha^i y^i \phi + \sum_i \xi^i (1 - \alpha^i) + \sum_i \alpha^i$$

$$\frac{\partial L}{\partial w} = 0 \qquad Cw = \sum_i \alpha^i y^i \phi$$

$\frac{\partial L}{\partial \xi}$: $\min_{\xi \geq 0} \xi^i(1-\alpha^i) \rightarrow \alpha \leq 1$

b/c otherwise $\alpha \rightarrow \infty$ as this would make $\xi^i(1-\alpha^i) \rightarrow -\infty$

$$g = \frac{C}{2}\left\|\frac{1}{C}\sum_i \alpha^i y^i \phi\right\|_2^2 - \frac{1}{C}\sum_i \alpha^i y^i \phi \sum_i \alpha^i y^i \phi$$

$$+ \sum_i \alpha_i$$

$$= \frac{C}{2}\cdot\frac{1}{C^2}\left\|\sum \alpha^i y^i \phi\right\|_2^2 - \frac{1}{C}\left\|\sum \alpha^i y^i \phi\right\|_2^2 + \sum \alpha_i$$

$\frac{1}{2C} - \frac{1}{C}$

$\frac{1}{2C} - \frac{2}{2C}$

$-\frac{1}{2C}$

$$= -\frac{1}{2}C\left\|\sum_i \alpha^i y^i \phi\right\|_2^2 + \sum \alpha_i$$

$$\max_{0 \le \alpha \le 1} g$$

Log Loss & Hinge loss:

$$\lim_{\epsilon \to 0} \epsilon \log\left(1 + \exp\frac{-F}{\epsilon}\right)$$

when $F \ge 0$ $= 0$

when $F \le 0$ $= -F$

$$\lim_{\epsilon \to 0} \epsilon \log\left(1 + \exp\frac{-F}{\epsilon}\right)$$

$$= \max(0, -F)$$

SVM is 0-temp limit of logistic regression

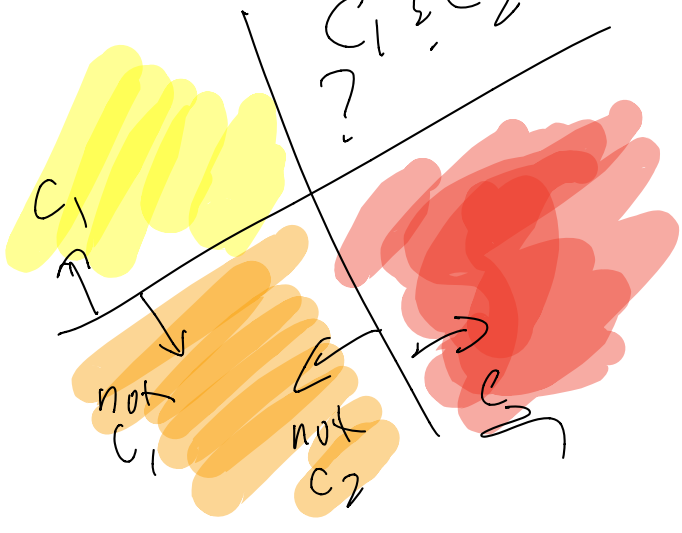
$$\min_w \frac{C}{2}\|w\|_2^2 + \sum \epsilon \log\left(1+\exp\left(\frac{L-yw\phi}{\epsilon}\right)\right)$$

$\epsilon \to 0 \Rightarrow$ SVM

$\epsilon = 1 \Rightarrow$ Log Reg.

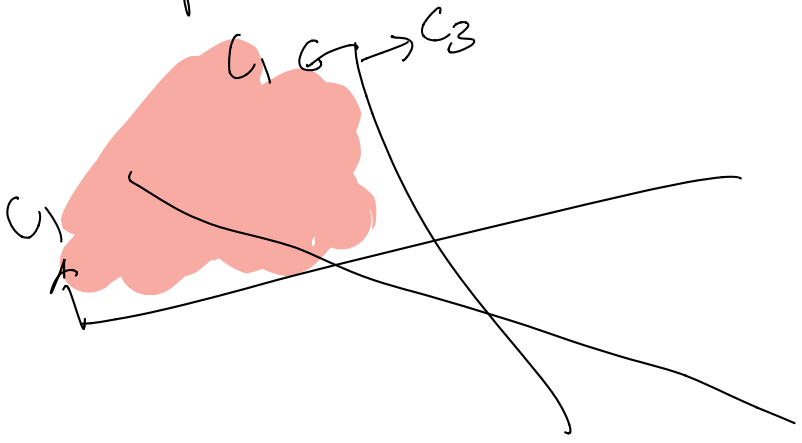
Multi Class

- 1 vs all
  - K-1 classifiers
  - is point in a class or not



- 1 vs. 1
  - $\frac{K(K-1)}{2}$, one for each

pair



- classify w/ majority vote

- Multi class Logistic Regression

$$P(y^{(i)} = k \mid x^{(i)}) = \frac{\exp w_{(k)}^T \phi(x^{(i)})}{\sum_{j \in \{0,\dots,K-1\}} \exp w_{(j)}^T \phi(x^{(i)})}$$

interpretal as probabilities

$$\arg\min_w \sum -\log P(y = y^{(i)} \mid x^{(i)})$$

minimize the sum of $-\log$ of the correct class probability

- Multi Class SVM

$$w_{y^{(i)}}^T \phi \geq w_{\hat{y}}^T \phi \quad \forall i \in D$$
$$\hat{y} \in \{0, \ldots, K-1\}$$

- we want correct class score ($w_{y^{(i)}}^T \phi$) to be bigger than all our scores for every other class

$$\min_w \frac{C}{2}\|w\|_2^2 + \sum_i \left(\max_{\hat{y}} \left(1 + w^T \psi(x^i, \hat{y})\right) - w^T \psi(x^i, y^i) \right)$$

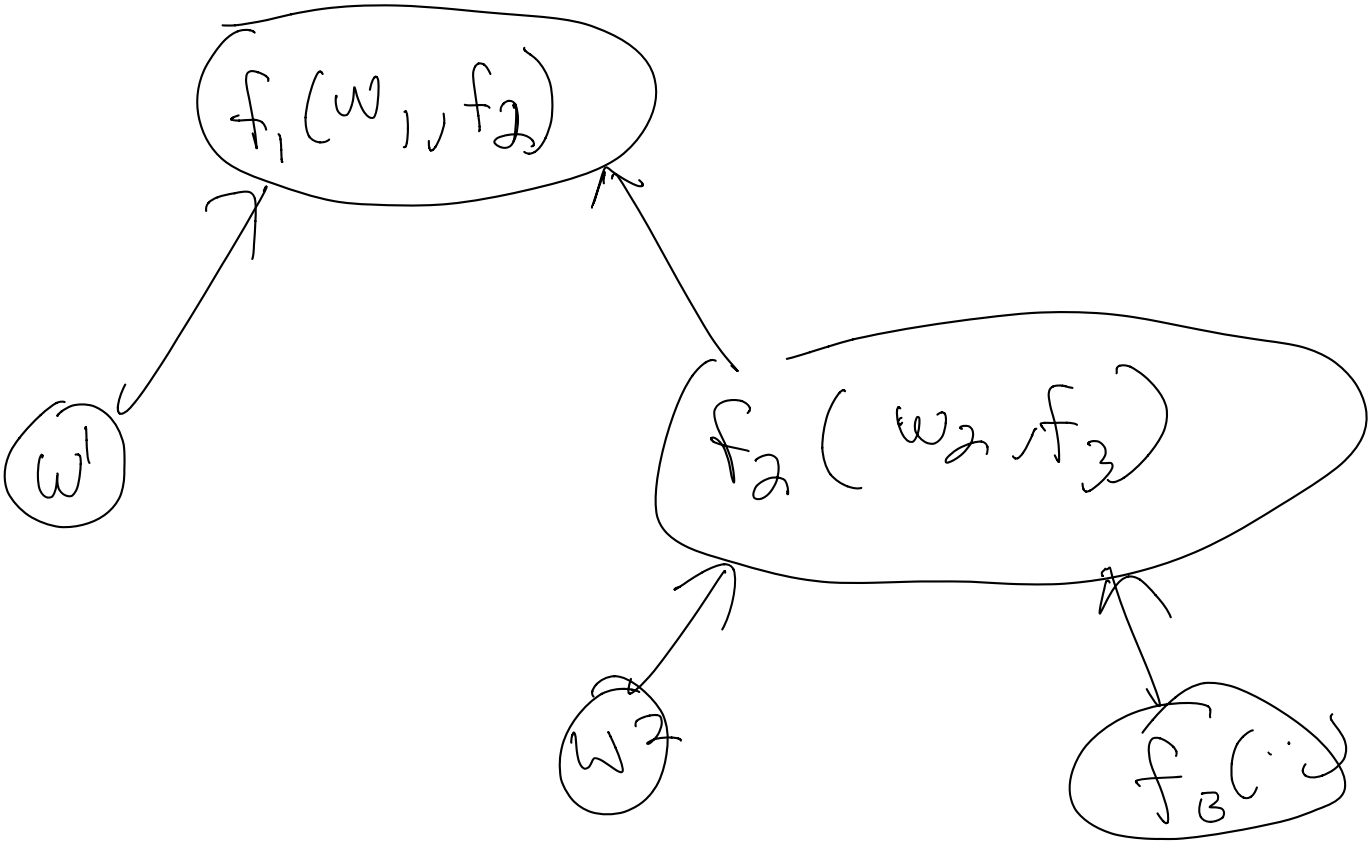
$\max_{\hat{y}} \left(1 + w^T \psi(x^i, \hat{y})\right)$: score our model gives (inference).

$w^T \psi(x^i, y^i)$: Score of correct class

- Generally

$$\min_w \frac{C}{2}\|w\|_2^2 + \sum_i \epsilon \ln \sum_{\hat{y}} \exp \frac{L(y^i, \hat{y}) + w^T \psi(x^i, \hat{y})}{\epsilon} - w^T \psi(x^i, y^i)$$

Deep Nets



- Conv Nets
  - decrease spatial resolution and # channels.
- Fully Connected

  $WX + b$

  - Trainable P's: w & b
  - issues b/c if $X \in 256 \times 256$, you will have ALOT of weights to tune.

- Convolutional Layer

  . decrease spatial resolution and # channels.

$W_1 \times H_1 \times D_1 \Rightarrow$ [CONV (K, F, S, P)] $\rightarrow W_2 \times H_2 \times D_2$

K = # filters
F = size
S = stride
P = padding

$$W_2 = \frac{W_1 - F + 2P}{S} + 1$$

$$h_2 = \frac{h_1 - F + 2P}{S} + 1$$

$$D_2 = K$$

$F \cdot F \cdot D_1$ weights / filter

$F \cdot F \cdot D_1 \cdot K$ weights total

- Max Pooling

  • reduce spatial size

  • prevent overfitting

* runs a window along an image and replaces the values in the window w/ max / average

$W_1 \times H_1 \times D_1 \Rightarrow$ [Pool F, S] $\Rightarrow W_2 \times H_2 \times D_1$

$$W_2 = \frac{W_1 - F}{S} + 1$$

$$H_2 = \frac{H_2 - F}{S} + 1$$

- Drop Out
  - randomly set activations to $\emptyset$
  - regularization
- Use NLL loss w/ softmax
- Train using Stochastic Gradient descent w/ momentum.

$$F(w_1, x, y) =$$

$$f_1(w_1, y, f_2(w_2, f_3(w_3, x)))$$

$$x_2 = f_3(w_3, x)$$

$$x_1 = f_2(w_2, x_2)$$

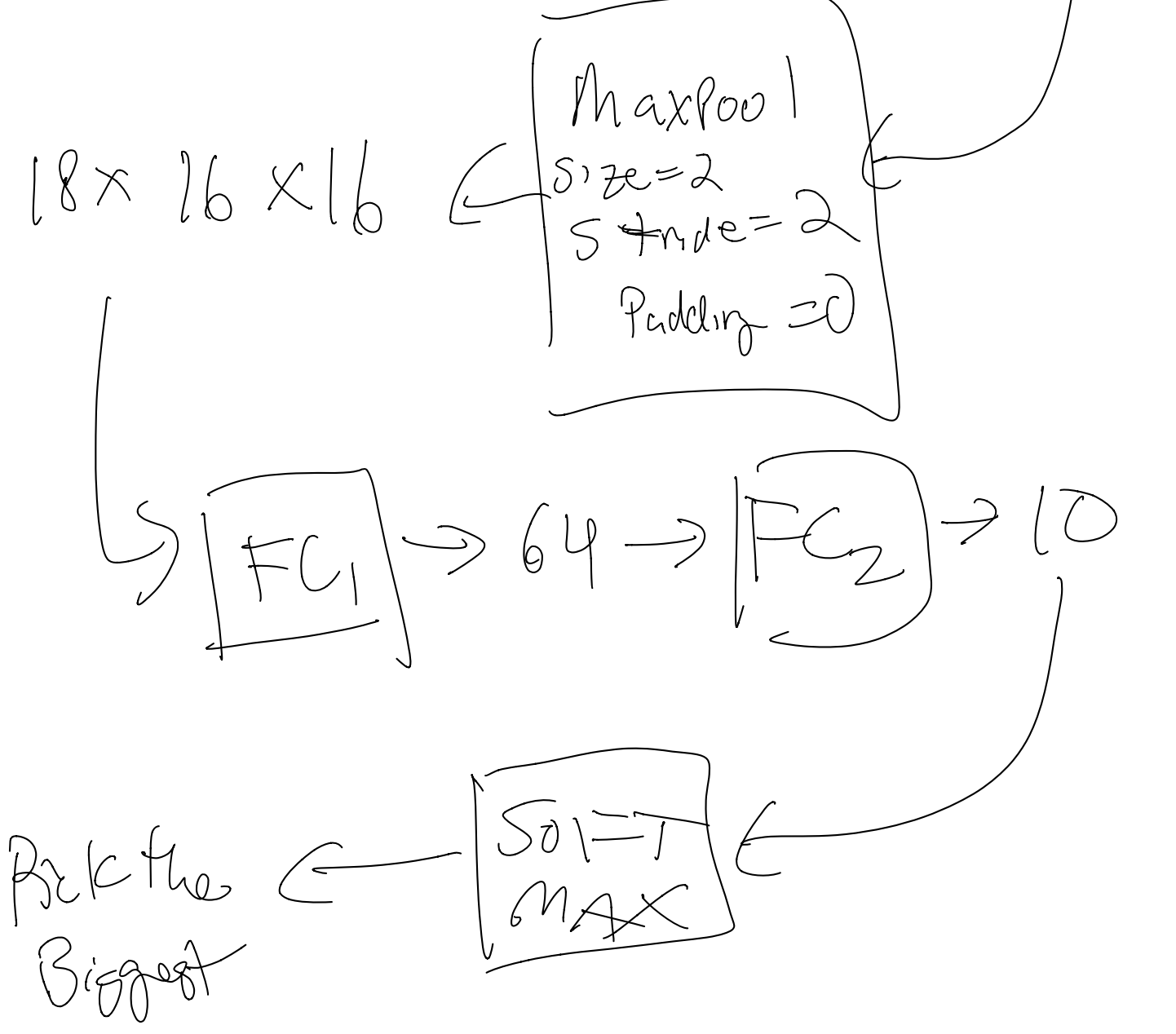
$$\frac{\partial F}{\partial w_3} = \frac{\partial f_1}{\partial x_1} \frac{\partial x_1}{\partial x_2} \cdot \frac{\partial x_2}{\partial w_3}$$

$$= \frac{\partial f_1}{\partial f_2} \cdot \frac{\partial f_2}{\partial f_3} \cdot \frac{\partial f_3}{\partial w_3}$$

$$\frac{\partial F}{\partial w_2} = \frac{\partial f_1}{\partial x_1} \cdot \frac{\partial x_1}{\partial w_2} = \frac{\partial f_1}{\partial f_2} \cdot \frac{\partial f_2}{\partial w_2}$$

$3 \times 32 \times 32 \rightarrow$ Conv1 $\rightarrow 18 \times 32 \times 32$

Conv1
18 filters
size=3
stride=1
Padding=1



Boosting

- weak classifiers working together to be a strong classifier
- Given:

$$D = \{(x^{(i)}, y^{(i)})\}$$

$F = \{f_t\}$ set of weak classifiers.

$$F_T(x) = \sum_{t=1}^{T} \alpha_t f_t(x)$$

$$F_t = F_{t-1}(x) + \alpha f_t(x)$$

Decision Tree

Inference

- Start at the root and follow decisions
- leaf node reveals results

Learning

- Chose a variable that "best" splits the data
- Split data according to rule
- Stop one # is small

Structured Prediction

- Inference
  - find the highest score from output
  - easy for binary - 1000 classes

QUIZ

Q U I Z → w/o structure

Q U I Z → w/ structure

- Formulate it as prediction of all 4 letter words.

$$\underline{26} \quad \underline{26} \quad \underline{26} \quad \underline{26} \rightarrow 26^4$$

feat.

- Size of feature space ≃ # classes

- Formally

$$Y = (y_1, \cdots, y_D) \quad y_d \in \{1, \cdots, K\}$$

$$Y^* = \underset{\hat{y}}{\operatorname{argmax}}\, F(w, x, \hat{y})$$

$$= \underset{\hat{y}}{\operatorname{argmax}}\, F(w, x, \hat{y}_1, \cdots, \hat{y}_D)$$

Basically, pick configuration w/ highest score