Example (regression $y \in \mathbb{R}$)

$\ell(v, v) = \frac{1}{2}(v - v)^2$ square loss / l_2

$\ell(v, v) = |v - v|$ absolute loss / l_1

$\ell(v, v) = |v - v|^p$ p-loss / l_p

If we take $\ell = l_2$ loss, then

$$R(f) = \mathbb{E}\, \ell(f(x), y) = \mathbb{E}\,(f(x) - y)^2$$

mean-squared prediction error of f (MSE)

$$= \mathbb{E}_X \left[\mathbb{E}_y \left[(f(x) - y)^2 \mid x \right] \right]$$

Now that

$$\mathbb{E}_\varphi[(f(x)-\varphi)^2 \mid X]$$

$$= \mathbb{E}_\varphi[f(x)^2 - 2f(x)\varphi + \varphi^2 \mid X]$$

$$= f(x)^2 - 2f(x)\underbrace{\mathbb{E}[\varphi \mid X]}_{\text{constant}} + \underbrace{\mathbb{E}_\varphi[\varphi^2 \mid X]}_{\text{constant}}$$

so this minimized when

$$2f(x) - 2\mathbb{E}[\varphi \mid X] = 0$$

$$\Rightarrow f(x) = \mathbb{E}[\varphi \mid X]$$

So the predictor f^* that minimizes the MSE risk in estimating Y is the conditional expectation of Y given X,

$$\mathbb{E}[Y|X]$$

f^* can be arbitrarily complicated, and we restrict ourselves to a particular function class, so

$$f^*_{\mathcal{F}} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \; R(f)$$

can be thought of as the best estimate of $\mathbb{E}[Y|X]$ in $\mathcal{F}$

Fact: if $\ell(u,v) = |u - v|$

then $f^* = \text{Med}(q|x)$

Recall we don't have access to $P_{\mathcal{D}}$
so we do ERM

$$\hat{f}_{\mathcal{F}} = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\ R_n(f)$$

$$= \underset{f \in \mathcal{F}}{\operatorname{argmin}}\ \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

and we hope

$$R(\hat{f}_{\mathcal{F}}) \approx R(f^*_{\mathcal{F}}) \approx R(f^*)$$

$\approx$: optimization + generalization error

$\approx$: approximation error: how good is $\mathcal{F}$

An additional concern is that we use numerical optimization, so we don't actually find $\hat{f}_{\mathcal{F}}$, instead we find $\tilde{f}_T$ say after T iterations of algorithm A.

Thus the error (risk) of models learned using ERM breaks down as

$$\mathcal{E}(A, n, \mathcal{F}, T) = R(\tilde{f}_T)$$

$$= \underbrace{R(\tilde{f}_T) - R(\hat{f}_{\mathcal{F}})}_{\text{optimiz error}} + \underbrace{R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}})}_{\text{generalization gap}}$$

$$+ \underbrace{R(f^*_{\mathcal{F}}) - R(f^*)}_{\text{approximation error}} + R(f^*)$$

can't get at

f^* — Bayes optimal estimator minimizes pop. risk

approximation error

$f^*_{\mathcal{F}}$ — best approx of f^* in our function class $\mathcal{F}$

generalization gap

$\hat{f}$ — best empirical risk minimizer in $\mathcal{F}$

optimization error

$\hat{f}_T$ — numerical minimizer of the empirical risk over $\mathcal{F}$

can get our hands on

Ex logistic regression (Maximum Likelihood Estimation give rise to ERM)

Assume we have predictor variables $x \in \mathbb{R}^d$ and a target $y \in \{\pm 1\}$, and we can model the log odds of $y=1$ linearly in x

$$\ln\left(\frac{p}{1-p}\right) = \underline{w_0} + \sum_{i=1}^{d} \underline{w_i x_i}$$

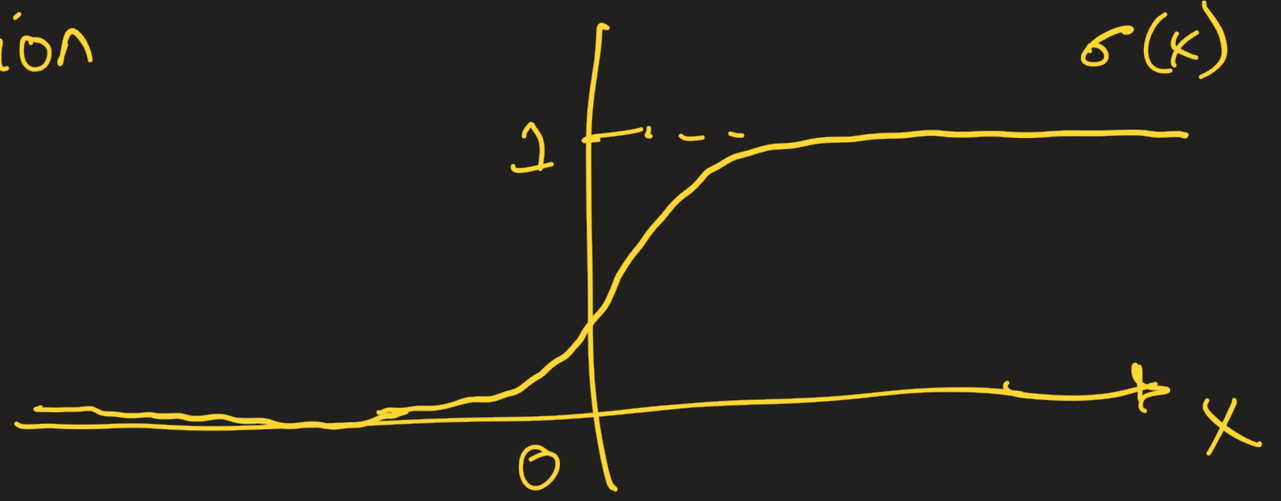
interpretation: probability of success goes up if $w_i > 0$ and x_i increases
goes down if $w_i < 0$ and x_i increases

$$\Rightarrow \frac{p}{1-p} = e^{\omega_0 + \omega^T x} \Rightarrow p = \frac{e^{\omega_0 + \omega^T x}}{1 + e^{\omega_0 + \omega^T x}}$$

$$\Rightarrow p = \frac{1}{\underbrace{1 + e^{-\omega_0 + \omega^T x}}}$$

$$p = \sigma(\omega_0 + \omega^T x)$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is called the sigmoid function



Now that means

$$y \sim \text{Bern}\left(\sigma(w_0 + w^T x)\right)$$

Goal: recover w_0 & w given training data
so in the future when we observe x
we can predict y.

Use maximum likelihood principle: choose the w_0, w that maximize the prob. of your training data.

$$w_0^*, w^* = \underset{w_0, w}{\operatorname{argmax}} \; p(y_1, \dots, y_n)$$

← prob. of my observed data if they were i.i.d. drawn from the model

$$= \underset{w_0, w}{\operatorname{argmax}} \; p(y_1, \dots, y_n)^{1/n}$$

$$y \sim \text{Bern}(\sigma(w_0 + w^T x))$$

$$= \underset{w_0, w}{\operatorname{argmax}} \left[\prod_{i=1}^{n} p(y_i) \right]^{1/n}$$

$$= \underset{w_0, w}{\operatorname{argmax}} \log\left(\dots \right)$$

$$= \underset{w_0, w}{\operatorname{argmax}} \sum_{i=1}^{n} \frac{1}{n} \log(p(y_i))$$

$$= \underset{w_0, w}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} -\log(p(y_i))$$

The quantity $-\log p(y_1, \ldots, y_n)$

$$= \sum_{i=1}^{n} -\log\left(p(y_i)\right)$$

is called the negative log-likelihood of the model parameters w_0, w

ML principle $\Rightarrow$ choose your model params by minimizing the negative log-likelihood

$$w_0^*, w^* = \underset{w_0, w}{\operatorname{argmin}} \; -\frac{1}{n} \sum_{i=1}^{n} \log(p(y_i))$$

$$p(y_i) = \begin{cases} \sigma(w_0 + w^T x_i) & \text{if } y_i = 1 \\ 1 - \sigma(w_0 + w^T x_i) & \text{if } y_i = -1 \end{cases}$$

fact: $\frac{1}{1+e^{-x}} + \frac{e^{-x}}{1+e^{-x}} = 1$

$\Rightarrow \boxed{1 - \sigma(x) = \sigma(-x)}$

$\sigma(x) \quad \frac{1}{1+e^{x}}$

$=$

$\sigma(-x)$

$\Rightarrow p(y_i) = \begin{cases} \sigma(w_0 + w^T x_i) & \text{if } y_i = 1 \\ \sigma(-(w_0 + w^T x_i)) & \text{if } y_i = -1 \end{cases}$

$$\Rightarrow \quad p(y_i) = \sigma\left(y_i\left(\omega_0 + \omega^T x_i\right)\right)$$

$$\Rightarrow \quad \omega_0^*, \omega^* = \operatorname{argmin} -\frac{1}{n} \sum \log\left(\sigma\left(y_i\left(\omega_0 + \omega^T x_i\right)\right)\right)$$

$$-\log\left(\sigma(x)\right) = \log\left(\frac{1}{\sigma(x)}\right) = \underbrace{\log\left(1 + e^{-x}\right)}_{\text{log loss!}}$$

$$\Rightarrow \quad \boxed{\omega_0^*, \omega^* = \operatorname{argmin} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-y_i\left(\omega_0 + \omega^T x_i\right)}\right)}$$

logistic regression

This is ERM w/ loss

$$\ell(v, v) = \log(1 + e^{-vv})$$

Note we took

$$y \sim \text{Bern}(\sigma(w_0 + w^T x))$$

we could say

$$y \sim \text{Bern}(\sigma(f(x)))$$

for $f \in \mathcal{F}$

then we get non-linear logistic regression

$$f^* = \underset{f \in \mathcal{F}}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-y_i \sigma(f(x_i))})$$

Next time:

- overfitting
- regularization (lasso, ridge regression)
- convex optimiz (poly regression)