MLE

Sheet 3

Exercise ①

① Under the joint distribution $X, Y \sim p(x,y)$ $\mathcal{L}(Y, f(X))$ is a r.v. Since $\mathcal{L}$ is bounded (and we'll assume integrable since any sensible choice of loss function is), $\mathcal{L}(Y, f(X)) = L$ has a defined expectation, $\mathbb{E}_{p(x,y)}(L)$. Since the draws y_i, x_i are iid, we have by LLN (more precisely, the strong law of large numbers)

$$\frac{1}{N} \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i)) \xrightarrow{a.s.} \mathbb{E}_{p(x,y)}(L)$$

substituting definitions,

$$R_{emp}(f, \mathcal{D}) \xrightarrow{a.s.} \mathbb{E}_{p(x,y)}(\mathcal{L}(Y, f(X))) = R(f).$$

As desired.

② ~~Suppose the distribution of training data does not match the true distribution that the training data is sampled from. This can happen very easily for small $|\mathcal{D}|$.~~

The training data consists of a finite sampling from a function (deterministic or random) and so there are (infinitely) many interpolations which achieve 0 loss on the finite data set, but do not match the true function (they can't all match since there's only one true function).

The bias - variance tradeoff states that ~~for a data set D, fixed x_0, and (potentially random) label distrib~~

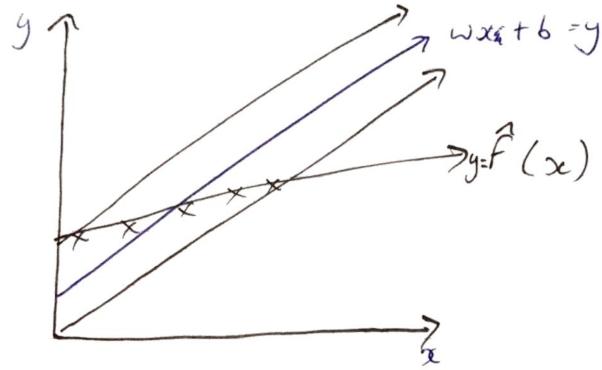
$$\mathbb{E}((Y - \hat{f}(x))^2) = \mathrm{Var}\, \hat{f}(x) + \mathrm{Bias}(\hat{f}(x))^2 + \sigma_e^2$$

when we randomly draw D, σ_e^2 is the variance of the error, supposing an additive noise model $Y = f(X) + \varepsilon$ with $\mathrm{Var}\, \varepsilon = \sigma_e^2$. We call this the irreducible error, which provides a minimum error (at least for this ~~model~~ error model and MSE loss).

This equation shows that it's not sufficient to minimise bias (which a very flexible model class permits), since such models have high variance under repeated draws of D. On the other hand an insufficiently flexible model class will have low variance under draws of D, but (at least on average) will have high bias. Thus overly- and underly- flexible model classes are to be avoided. The lessons generalise beyond MSE loss and additive error.

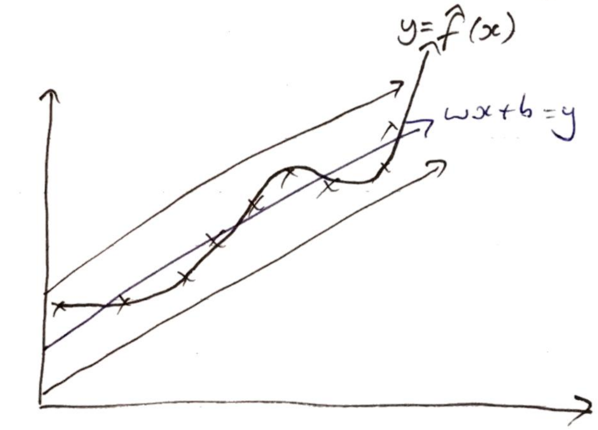
(b) Blue • = $wx_i + b$, Black • = 95% error bar bounds. x = data draws.

$\mathcal{H}_1$



$\hat{f}$ achieves low empirical error, but doesn't match ideal fit — well: high R.

$\mathcal{H}_2$



c) K-fold CV simulates K draws from D, allowing an estimate of $\mathrm{Var}_{\mathcal{D}}\mathcal{L}(f_m, Y)$, allowing us access to another component within the bias-variance tradeoff.

③ a) By the application of the tower law in the hint,

$$R(f) = \mathbb{E}_{p(x)}\left[R(f|\underline{x})\right] = \mathbb{E}_{p(\underline{x})}\left[\underbrace{\mathbb{1}(f(\underline{x}) \neq y \mid X = \underline{x})}_{\text{No longer random}}\right]$$

So ~~to~~ minimise $R(f)$, it suffices to minimise $\mathbb{1}(f(\underline{x}) \neq y \mid X = \underline{x})$ for all choices of $\underline{x}$. But we must have $f(\underline{x}) \in \{0,1\}$ to achieve this minimality since $y \in \{0,1\}$, and minimisation over ~~$p(\underline{x})$~~ y exactly is achieved when the more probable y is chosen:

$$f(x) = \underset{k \in \{0,1\}}{\operatorname{argmax}}\ p(y = k \mid \underline{x})$$

which is exactly the MAP rule.

b) Optimal choice of $\hat{f}$: we wish to minimise

$$R(f) = \mathbb{E}_{p(\underline{x})}\left[R(f|\underline{x})\right]$$

$$= \mathbb{E}_{p(x)}\left[\mathbb{E}_{p(y)} \|y - f(\underline{x})\|_2^2 \mid X = \underline{x}\right].$$

so for each choice of $\underline{x}$, we wish to ~~wish~~ choose $\hat{y}_{MAP}$ that minimises $\| y - \hat{y}_{MAP}\|_2^2$ over $y \sim p(y)$. It is known that the minimiser of the squared L_2 norm is the mean. Thus choose

$$\hat{f}_{MAP} : \underline{x} \longmapsto \mathbb{E}(Y \mid X = \underline{x})$$

4 a) Notice, as before, that $L = \mathcal{L}(y, f(\underline{x}))$ is a r.v. depending on draws from $y \sim p(y)$ and $f(\underline{x})$, $x \sim p(x)$. Since $R(f) = \mathbb{E}(L)$ and $R_{emp}(f|D) = \frac{1}{N}\sum_{i=1}^{N} L_i$ with $L_i = \mathcal{L}(y_i, f(\underline{x}_i))$, we can apply Hoeffding's bound since (for fixed f) $\mathcal{L}$ is bounded.

Let $\Delta = |R_{emp}(f|D) - R(f)|$, and $d = \sqrt{\frac{M^2 \ln(2/\delta)}{2N}}$.
Then

$$\Pr(\Delta < d) \geq 1 - \delta.$$

$$\Leftrightarrow \quad \delta \geq 1 - \Pr(\Delta < d) = \Pr(\Delta \geq d)$$

Now applying Hoeffding's bound with $\varepsilon = d$,

$$\Pr(\Delta \geq d) \leq 2\exp\left(-\frac{2N}{M^2} d^2\right)$$

$$= 2\exp\left(-\frac{2N}{M^2} \frac{M^2 \ln\frac{2}{\delta}}{2N}\right)$$

$$= 2\exp\left(-\ln\frac{2}{\delta}\right) = 2\exp\left(\ln\frac{\delta}{2}\right)$$

$$= \delta$$

So $\Pr\left(\Delta \geq \sqrt{\frac{M^2 \ln(2/\delta)}{2N}}\right) \leq \delta$

$\Rightarrow \Pr\left(\Delta < \sqrt{\frac{M^2 \ln(2/\delta)}{2N}}\right) \geq 1 - \delta$, as desired.

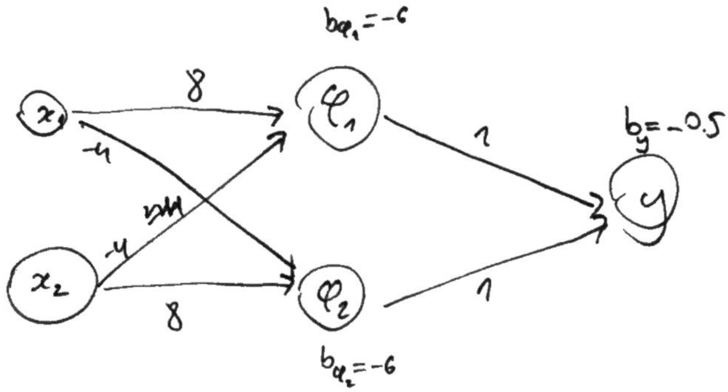
b)

b) We want $\Delta \approx 0$ and we have finite N. The bound goes as $\frac{1}{\sqrt{N}}$, so increasing by a factor c the number of samples decreases the (probabilistic) bound by $\frac{1}{\sqrt{c}}$. On the other hand, the bound goes with μ, so it's much more~~a~~ (quadratically in fact) effective if one can reduce the loss bound, rather than the number of samples.

We can also write $\sqrt{\frac{\mu^2 \ln(\frac{2}{\delta})}{2N}} \approx 1.92 \frac{\mu}{\sqrt{2N}}$ when $\delta = 0.05$ (ie ~~when~~ the bound holds 95% of the time). A factor of only ~ 2 isn't so bad, and it means that the μ (which isn't much under our control) ~~is~~ t can be the main factor.

Exercise 2

(1)



(I'm not sure where the biases should be drawn, as it sounds like they're also supposed to have arrows, but the usual drawing ~~ot~~ of a network doesn't easily permit that. x_1, x_2 are drawn in circles even though no activation is applied to them.)

Exercise 2: From Logistic Regression to Neural Networks

4.) For such a network with the identity as the activation function the network output is:

$$z^{(L)} = W^{(L)} \left(W^{(L-1)}\left(\dots \left(W^{(2)}\left(W^{(1)} x + b^{(1)}\right) + b^{(2)}\right) + \dots + b^{(L-1)}\right) + b^{(L)}\right)$$

This can be rewritten using:

$$\tilde{b} = b^{(L)} + W^{(L)} b^{(L-1)} + W^{(L)} W^{(L-1)} b^{(L-2)} + \dots + W^{(L)} \dots W^{(2)} b^{(1)} = \sum_{l=1}^{L} \left(\prod_{k=1}^{L-l} W^{(L-k)} \right) b^{(l)}$$

$$\tilde{W} = W^{(L)} W^{(L-1)} \dots W^{(1)} x = \prod_{l=1}^{L} W^{(L-l)} x$$

Yielding

$$z = \tilde{W} x + \tilde{b}$$

a linear 1-layer network.

In order to extend the expressiveness of the network beyond linear transformation, non-linearities such as sigmoid activation are required, since however intricate the layer structure might be, if only linear operations are performed within it, it can always be reduced to a simple 1-layer net.