Bayesian Deep Learning

Jan 9, 2021

MASH & IASD

Acknowledgement:

I've borrowed many ideas from Andrew Gordon Wilson, NYU.

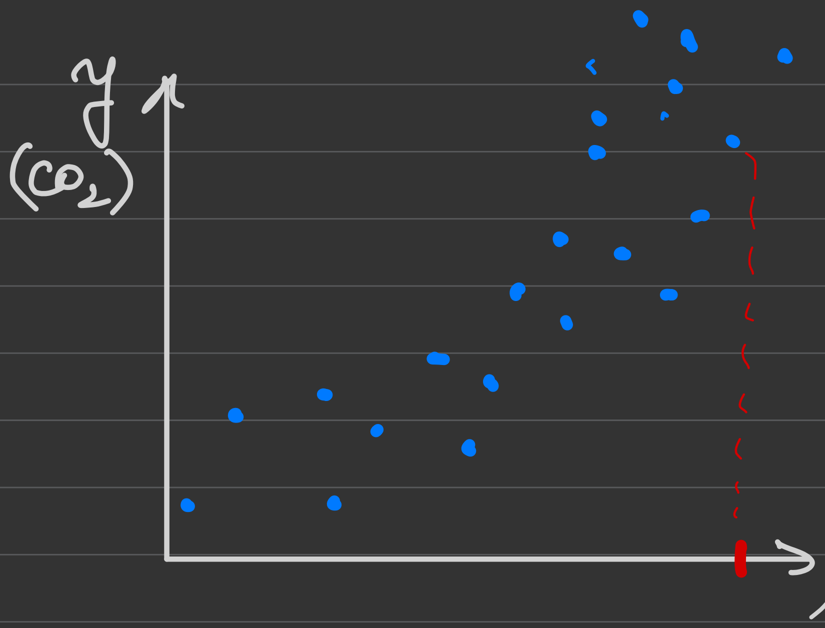
What comes to your mind when you hear

Bayesian inference?

probability distribution from data

update prob. given initial belief

Deep Learning



(1) $y(t) = a_0 + a_1 t$

(2) $y(t) = \sum_{i=0}^{3} a_i t^i$

(3) $y(t) = \sum_{i=0}^{10^4} a_i t^i$

Uncertainty:

- when gathering data: aleatoric uncertainty
  → irreducible uncertainty
- choice of model / hypothesis
  → epistemic uncertainty
  → reducible uncertainty

$$f(x;w) = \sum_{j=0}^{J} w_j x^j$$

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1\ldots n}$$

learn $\underline{w} = (w_0, w_1, \ldots, w_J)$

$$\mathcal{L}_{SE}(\underline{w}) = \sum_{i=1}^{n} \left(f(x_i;\underline{w}) - y_i \right)^2$$

minimize wrt $\underline{w}$

$$\mathcal{L}_{AE}(\underline{w}) = \sum_{i=1}^{n} \left| f(x_i;\underline{w}) - y_i \right|$$

$$y(x) = f(x; \underline{w}) + \underbrace{\varepsilon(x)}_{} \quad \text{noise}$$

$$y(x) \sim N(f(x, \underline{w}), \sigma^2) \Longleftrightarrow \varepsilon(x) = \varepsilon \sim N(0, \sigma^2)$$

Likelihood $\mathcal{D} = \{(x_i, y_i)\}_{i=1 \dots n}$

$$p(y_1, \dots, y_n \mid x_1, \dots, y_n, \underline{w})$$

$$= \prod_{i=1}^{n} p(y_i \mid x_i, \underline{w}) \quad \text{(by conditional iid)}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(f(x_i, \underline{w}) - y_i)^2\right)$$

$$\text{log-like} = \log\left[\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\right] - \underbrace{\frac{1}{2\sigma^2}\sum_{i=1}^{n}(f(x_i, \underline{w}) - y_i)^2}_{\text{Same as previous objective.}}$$

☒ Gaussian noise $\rightsquigarrow$ squared error.

☒ What about other distr. for noise

e.g. Laplace : $p(w) = \frac{1}{2} e^{-|x|}$

heavier than Gaussian

Bayesian modeling

Like. $p(y \mid x, \underline{w}) = N(y; f(x; \underline{w}), \sigma^2)$

Prior. $p(\underline{w}) = N(0, \gamma^2 I)$

$\iff \forall j, \; p(w_j) = N(0, \gamma^2)$

Sum rule $p(a) = \sum_b p(a, b)$

Product rule $p(a, b) = p(a)\, p(b \mid a)$
$= p(b)\, p(a \mid b)$

$\Rightarrow$ Posterior: $p(\underline{w} \mid \mathcal{D}) = \dfrac{\overbrace{p(\underline{w})}^{\text{prior}}\; \overbrace{p(\mathcal{D} \mid \underline{w})}^{\text{like.}}}{\underbrace{p(\mathcal{D})}_{\substack{\text{marginal likeli.} \\ \text{evidence}}}}$

Take log:

$\log p(\underline{w} \mid \mathcal{D}) = \log p(\underline{w}) + \log p(\mathcal{D} \mid \underline{w}) - \underbrace{\log p(\mathcal{D})}_{\text{not dep on } \underline{w}}$

Penalty $= \log$ prior

$= -\dfrac{1}{2\gamma^2}\, \underline{w}^T \underline{w}$

MAP
|| Maximum a posteriori

L^2 penalty, aka weight decay in ML/DL literature.

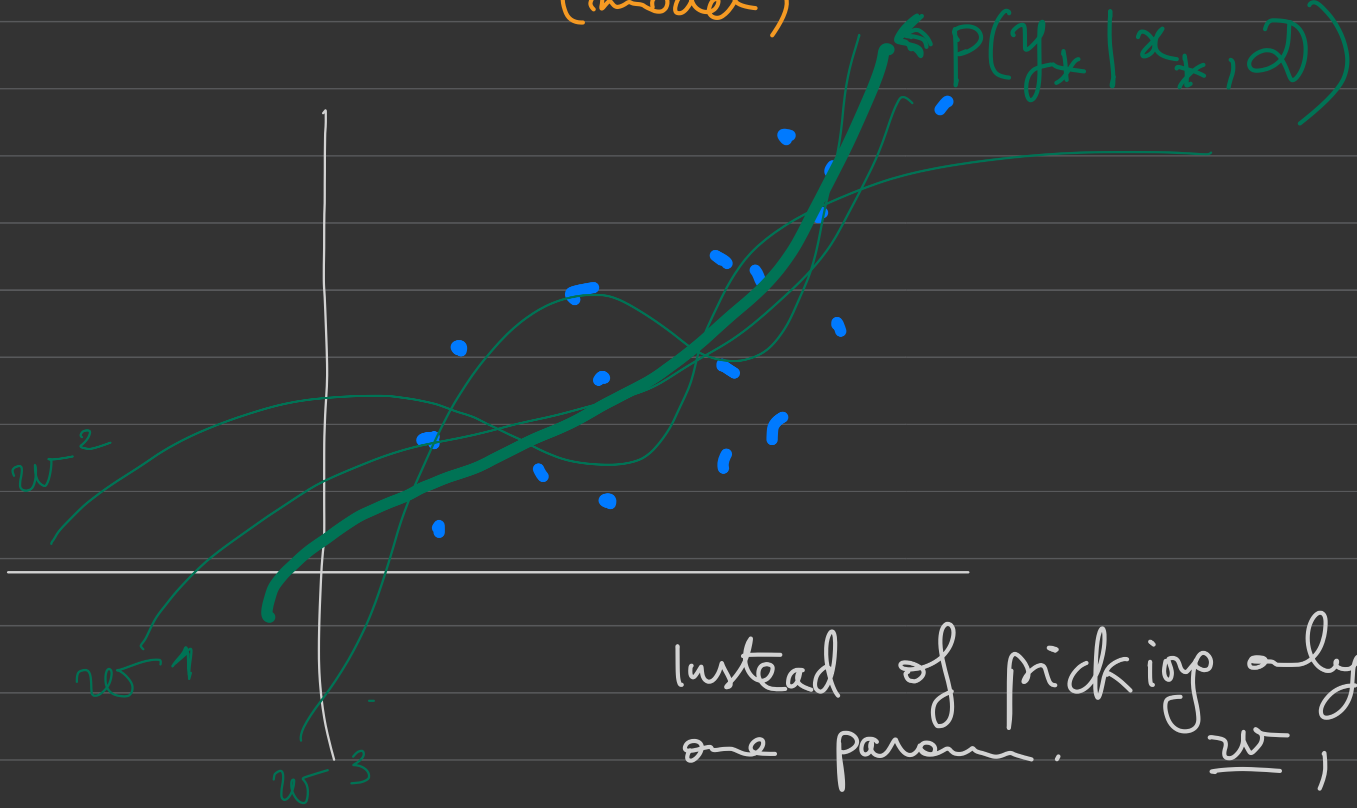
MAP is not quite Bayes because it's optim°

Instead: we want a full predictive distr.

$$p(y_* \mid x_*, \mathcal{D}) = \int p(y_*, \underline{w} \mid x_*, \mathcal{D})\, d\underline{w}$$

$$= \int p(y_* \mid \underline{w}, x_*, \cancel{\mathcal{D}})\, p(\underline{w} \mid \cancel{x_*}, \mathcal{D})\, d\underline{w}$$

$$= \int \underbrace{p(y_* \mid \underline{w}, x_*)}_{\text{likelihood (model)}}\, \underbrace{p(\underline{w} \mid \mathcal{D})}_{\text{posterior}}\, d\underline{w}$$



Instead of picking only one param. $\underline{w}$,

Bayesian model averaging takes them all and weight them w.r.t. posterior.

Quick detour:

1) Likelihood of flipping n coins w/ prob. λ of tail.

$y_i = 1$ when tail

$$p(y_1, \dots, y_n \mid \lambda) = \prod_{i=1}^{n} p(y_i \mid \lambda)$$
$$= \lambda^{\sum y_i} (1-\lambda)^{n - \sum y_i}$$

Binomial $(m \mid n, \lambda)$

$$p(m \mid n, \lambda) = \binom{n}{m} \lambda^{m} (1-\lambda)^{n-m}$$

2) MLE for λ?

$$\hat{\lambda}^{MLE} = \operatorname{argmax} p(m \mid n, \lambda) = \frac{m}{n} = \overline{Y}$$

3) Suppose you observed one tail.

What's the prob. that the next flip is tail again? (using MLE).

$$m = 1, \; n = 1 \Rightarrow \hat{\lambda}^{MLE} = \frac{1}{1} = 1$$

Bayes. approach

like $p(m \mid n, \lambda)$

prior $p(\lambda) = \text{Beta}(\lambda; a, b)$ $a, b > 0$

$\Rightarrow$ conjugate, so

post. $p(\lambda \mid m, n) = \text{Beta}(\lambda; a+m, b+n-m)$

$$\hat{\lambda}^{\text{Bayes}} = \mathbb{E}[\lambda \mid m, n] = \boxed{\frac{a+m}{a+b+n}}$$

$a = b$ $\hat{\lambda}^{\text{Bayes}} < 1$

Predictive distr.:

$$p(y_* \mid x_*, \mathcal{D}) = \int p(y_* \mid \underline{w}, x_*)\, p(\underline{w} \mid \mathcal{D})\, d\underline{w}$$

MLE chooses only one specif $\underline{w}^0$, ie chooses an approx. post $\approx \delta(\underline{w}^0)$.