DD2412 - Deep Learning Advanced Course

Labelled data $D = \{(\bar{X}_i, y_i)\}$ $i \in 1:n$

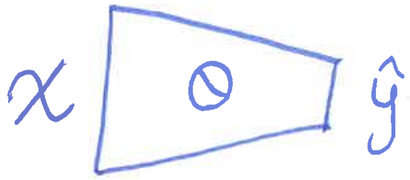
Input Label

classification $y \in 1, 2, 3 \dots k$

Regression $y \in \mathbb{R}$

weights $\theta \in \Theta$

$f_\theta : x \mapsto \hat{y}$



$\theta^* = \arg\min_\theta L(D)$

Find θ where the loss function over all D is minimal

label

$$L(D) = \sum_{i=1}^{n} l(f_\theta(x_i), y_i) + \Omega(\theta)$$

Regularization

eg: $l = MSE$

$\Omega = \|\theta\|^2 \; (L_2)$

Probabilistic descriminative learning

$\theta^* = \arg\max_\theta P(\theta | D)$

What parameters are most likely to match D

$$\Rightarrow \arg\max \frac{P(D|\theta) P(\theta)}{\cancel{P(D)}}$$

$$\arg\max \prod_i P(x_i, y_i | \theta) \, P(\theta)$$

$$\arg\max \sum \underbrace{\log(P(x_i, y_i | \theta))}_{P(y_i | x_i, \theta)} + \log(P(\theta))$$

Bayesian Modeling

$$P(\theta | D) = \frac{P(D|\theta) P(\theta)}{P(D)}$$

Posterior Likelihood evidence Prior

$\theta^* = \arg\max P(D|\theta) P(\theta)$

Usually the evidence is hard to calculate so it needs to be approximated

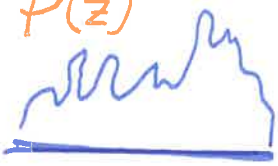
Can be used to make new likelyhood predictions of new datapoints

Variational Inference

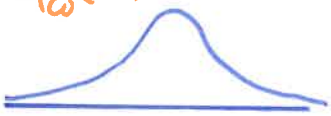
General Steps

- we want to find a distribution P (eg $P(\Theta|D)$)
- but the distribution is hard to find directly
- Settle with some approximation of P from a simpler family of distributions, Q parametarized by ω
- Find a parameter ω^* that minimizes difference between Q and P (KL divergence)

$P(z)$



$Q_\omega(z)$



find a ω^* that is as similar to P as possible
↓
Lowest KL-divergence

KL-divergence

$D_{KL}(P\|D)$ or $KL(P\|D)$ measure to compare P to D

$KL(P\|D)=0 \iff P=D$

$$KL(P\|D) = E_p\left[I_Q(x) - I_p(x)\right]$$
$$= \sum\left(-\log Q(x) + \log P(x)\right)p(x)$$
$$= \sum P(x)\log\frac{P(x)}{Q(x)}$$
$$= H(P,Q) - H(P)$$

$KL(P\|D) \Rightarrow$ Mean matching (moment)

$KL(D\|P) \Rightarrow$ Mode matching

↳ usually the one used

Reminders:

Information $I_p(X) = -\log P(X)$

Entropy $H(P) = E_p[I_p(x)] = -\sum \log(p)\, p(x)$

↳ 0 ⇒ 100% predictible

↳ max ⇒ uniform P

three observations.

KL is not symmetric.

for some $z \in P$

- $P(z)$ and $Q(z)$ is high ⇒ KL low
- $P(z)$ is low $Q(z)$ is high ⇒ KL still low
- $P(z)$ is high $Q(z)$ low ⇒ KL high

⇒ Q must be high whenever p is but not neccesarily other way around

Steps of Variational Inference

given observations X and hidden variables Z and we are interested in the true posterior. $P(Z|x)$

Use bayes rule $P(Z|x) = \dfrac{P(x|Z)P(Z)}{P(x)}$

Consider a simpler distribution $Q_\omega(Z)$, match Q to $P(Z|x)$ ↳ $P(x)$ is hard to find

$$P(Z|x): \min_\omega D_{KL}(Q_\omega(Z) \| P(Z|x))$$

↳ optimization instead of integration.
↳ Q is diferentiabe

Discriminative Bayesian modeling using VA

1. have training data $D = \{(x_1, y_1), \ldots (x_n, y_n)\}$
2. Assume prior distribution $P(\theta)$
3. Design Likelyhood $P(y|x, \theta)$
4. Decide on approximation distribution $Q_\omega(\theta)$
5. Find ω with the following objective

$$\max_\omega \sum \left(Q_\omega(\theta) \sum_i^n \log P(y_i | x_i, \theta) \right) - D_{KL}(Q_\omega(\theta) \| P(\theta))$$

Predictive Uncertanty.

We want the model to give out a certanty with its prediction both classification & regression

Datapoints that the model is uncertain about are the ones where it should focus on more

Why?

Small dataset, noisy in/out-put, incomplete input, covariate shift

Epistemic Vs Aleatoric

↳ Lack of Knowledge
- not enough data
- Networks ~~too~~ simple
- Bad optimizer
- Not same distribution

↳ Model Uncertanty

↳ Distributional uncertanty

↳ Unobtainable Knowledge
- Label noise
- Measurement precision
- ~~measurement~~ noise
- Class definition overlap

↳ Homoscedastic Uncertanty
- Absolute uncertainty
- constant uncertanty

↳ Hetroscedastic Uncertanty
- Relative uncertainty
- input dependent

How to model predictive uncertanty?

How to model Aleatoric Uncertanty?

We dont know the noise in the data

$$y = y_{true} + \text{noise}$$

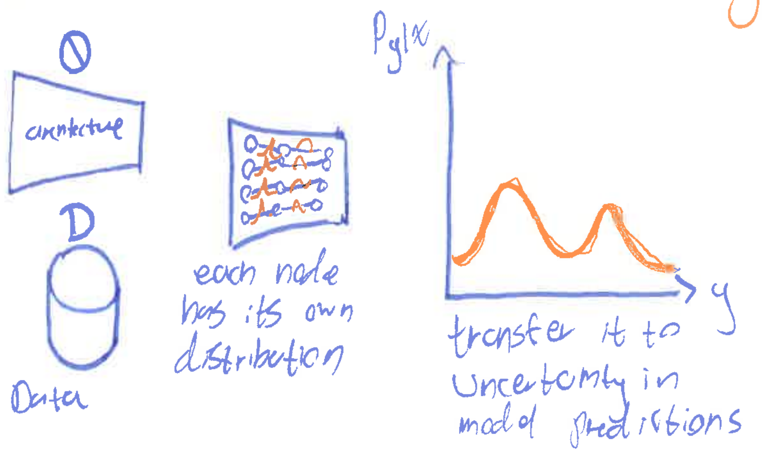
↳ ε independent of x ? ε vs $\varepsilon(x)$ $\quad \mathcal{N}(0, \sigma^2)$ $\quad \mathcal{N}(0, \sigma(x)^2)$

Assume a distribution (typpically gaußian)

1. have the model f_θ output a distribution over y
2. find θ with MLE
3. have loss according to

$$l(f_\theta(X_i), y_i) = \frac{\log a^2}{2} + \frac{(f_\theta(x_i) - y_i)^2}{2a^2} + C$$

How to model epistemic Uncertainty



Frequentist approach

train model multiple times and model the amount of disagreement between models.

Bayesian approach

$$P(\Theta|D) = \frac{P(D|\Theta)\,P(\Theta)}{P(D)}$$

get $P(y|\underline{x}, D)$

for discriminative

As in MAP
assume a prior $P(\Theta)$
devise a likelyhood function $p(y|x,\Theta)$

assuming D is iid
full posterior over Θ $P(\Theta|D)$: $P(\Theta|D) = \frac{\prod_i P(y_i|x_i,\Theta)\,P(\Theta)}{P(D)}$

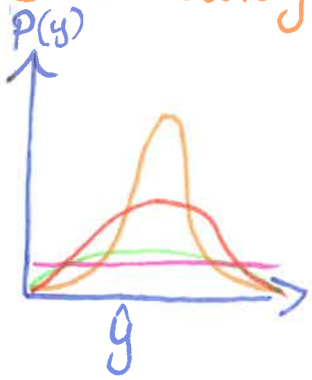
Varientional Inferense for Deep uncertainty Estimation

↳ Example

$* =$

1 have data D
2 have prior $P(\Theta) = \mathcal{N}(0, I)$
3 Design likelyhood $P(y|x,\Theta) = f_{NN}(x;\Theta)$ -softmax
4 asign approximatin function $Q_\omega(\Theta) = \mathcal{N}(\mu, \sigma^2 I)$
5 Find ω that maxemises *

Uncertainty Estimation



Aleatoric Uncertainty → Maximum Likelihood estimation of noise

irreducible

Epistemic Uncertainty

Decreases with data

Bayesian Modeling

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

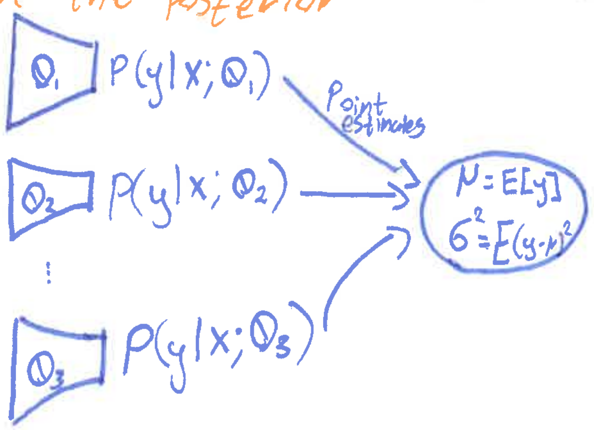
Variational approximation

$$\min_{\omega} D_{KL}(Q_{\omega}(z) \| P(z|x))$$

Ensemble Methods

Train S independent networks with bootstrapped data.

Have each network act as samples of the posterior



- \+ Not as many parameters
- \+ Model structure is unchanged
- \+ Parallelizable
- \- evaluation is inefficient
- \- lots of memory

SWAG

take snapshots of the model at different steps of the training and treat each estimate as its own network

Distilation

train an ensemble of networks, use them to label data. Train a student network on the larger dataset.

Evaluating Uncertainty

Proper Scoring

Imagine a scoring function $S(P_\theta(y|x), (x_i, y_i))$ that measures the quality of a predictive distribution with underlying data.

Define $S(P_\theta, P) = \iint P(x,y) S(P_\theta, (x,y))\, dx\, dy$

S is the proper scoring if

$$S(P_\theta, P) \leq S(P, P)$$

S is strictly proper iff

$$P_\theta(y|x) = P(y|x)$$

Log likelihood

$$S(P_\theta, (x_i, y_i)) = \log(P_\theta(y = y_i | x_i))$$

Brier score

$$S(P_\theta, (x_i, y_i)) = MSE(P_\theta(y|x_i), \delta_y)$$

Calibration

Measure statistical consistency between predictive and observed distribution

Log likelihood

Out of Distribution Detection

What should the model do when test and training data come from different distributions. Possibility is to abstain from prediction.