Lecture 14: Bayesian regression

Bayesian Inference

- Idea
  - Weights with a better fit to the training data should be more probable than others
  - Make predictions with all these weightsm scaled by their probability
- Reason under all possible parameter values
  - weighted by their posterior probability
- More robust predictions
  - less sensitive to overfitting, particularly with small training sets
  - Can give rise to more expensive model class

Frequentist v.s. Bayesian

- Frequentist: **learning using point estimates**, regularisation, p-values
  - backed by sophisticated theory in simplifying assumptions
  - mostly simpler algorithms, characterises much practical machine learning research
- Bayesian: maintain **uncertainty**, marginalise out unknowns during inference
  - some theory
  - often more complex algorithms, but not always
  - often more computationally expensive

Bayesian Regression

- Application of bayesian inference to linear regression, using normal prior over w
- Consider full posterior $p(w|X, y, \sigma^2)$
- Sequential Bayesian updating
  - Can formula $p(w|X, y, \sigma^2)$ for given dataset
  - As we see more and more data:
    1. Start with prior $p(w)$
    2. See new labelled datapoint
    3. Compute posterior $p(w|X, y, \sigma^2)$
    4. The **posterior now takes role of prior** & repeat from step 2

Conjugate Prior

- Product of **likelihood** $\times$ **prior**: results in the same distribution as the prior

Stages of Training

1. Decide on model formulation & prior
2. Compute **posterior** over parameters $p(w|x, y)$
3. 3 methods:
   1. MAP:
      1. Find mode for w
      2. Use to make prediction on test
   2. Approx. Bayes:

1. Sample many w
2. Use to make ensemble average prediction on test

3. Exactly Bayes
    1. Use all w to make expected prediction on test

Prediction with uncertain w

- Could predict using sampled regression curves
    - Sample S parameters, $w^{(s)}, s \in \{1, ..., S\}$
    - For each sample, compute prediction $y_*^{(s)}$ at test point x_*
    - (Monte Carlo integration)
- For Bayesian regression, there's a simpler solution:
    - Integration can be done analytically, for
    - $p(\hat{y}_*|X, y, x_*, \sigma^2) = \int p(w|X, y, \sigma^2)p(y_*|x_*, w, \sigma^2)dw$
- Pleasant properties of Gaussian distribution means integration is tractable
    - $p(\hat{y}_*|X, y, x_*, \sigma^2) = ... = \mathrm{Normal}(y_*|x_*'w_N, \sigma_N^2(x_*))$
    - $\sigma_N^2 = \sigma^2 + x_*'V_N x_*$
    - Additive variance based on x_* match to training data

Caveats (Notes)

- Assumption
    - known data noise parameter $
    - sigma^2$
    - data was drawn from the model distribution

Lecture 15: Bayesian classification

Discrete Conjugate prior

- Example:
    - Prior: Beta
    - Likelihood: Binomial
    - Posterior: Beta (conjugacy)

Suite of useful conjugate priors

- Regression:
    1. For mean:
        - Likelihood: Normal
        - Prior: Normal
    2. For variance / covariance:
        - Likelihood: Normal
        - Prior: Inverse Gamma / Inverse Wishart
- Classification:
    1. Likelihood: Binomial, Prior: Beta
    2. Likelihood: Multinomial, Prior: Dirichlet
- Counts:

1. Likelihood: Poisson, Prior: Gamma

Bayesian Logistic Regression

- Discriminative classifier which conditions on inputs
- Similar problems with parameter uncertainty compared to regression
- Need prior over w (coefficients), not q
- **No known conjugacy**
    - Thus, use a Gaussian prior
- Resolve by (Laplace) approximiation:
    - Assume posterior $\approx$ Normal about mode
    - Can compute normalisation constant, draw samples, etc.