COMP90051

Workshop Week 10

About the Workshops

- 7 sessions in total
  - Tue 12:00-13:00 AH211
  - Tue 12:00-13:00 AH108 *
  - Tue 13:00-14:00 AH210
  - Tue 16:15-17:15 AH109
  - Tue 17:15-18:15 AH236 *
  - Tue 18:15-19:15 AH236 *
  - Fri 14:15-15:15 AH211

About the Workshops

- Homepage
  - https://trevorcohn.github.io/comp90051-2017/workshops
- Solutions will be released on next Friday (a week later).

Reminder

- Project 2
  - Kaggle competition due on Mon, 09/Oct/17
  - Worksheet, report, and code due on Wed, 11/Oct/17

- Exam
  - Fri, 03/Nov/2017, 8:30am
  - 3 hours
  - Royal Exhibition Building

Syllabus

1	Introduction; Probability theory	Probabilistic models; Parameter fitting	
2	Linear regression; Intro to regularization	Logistic regression; Basis expansion	
3	Optimization; Regularization	Perceptron	
4	Backpropagation	CNNs; Auto-encoders	
5	Hard-margin SVMs	Soft-margin SVMs	
6	Kernel methods	Ensemble Learning	
7	Clustering	EM algorithm	
8	Principal component analysis; Multidimensional Scaling	Manifold Learning; Spectral clustering	
9	Bayesian inference (uncertainty, updating)	Bayesian inference (conjugate priors)	←
10	PGMs, fundamentals	PGMs, independence	
11	Guest lecture (TBC)	PGMs, inference	
12	PGMs, statistical inference	Subject review	

Outline

- Review the lecture, background knowledge, etc.
  - MLE, MAP, Bayesian estimates
  - Comparison between Bayesian and frequentist
  - Likelihood, prior, and posterior
  - Conjugate prior and likelihood
    - Bayesian linear regression

  - IPython notebook task: Bayesian linear regression

MLE, MAP

- Training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, $\boldsymbol{X}$ for all $\boldsymbol{x}_i$, $\boldsymbol{y}$ for all y_i
- $\widehat{\boldsymbol{w}} = \max_{\boldsymbol{w}} \prod_{i=1}^{N} p(y_i | \boldsymbol{x}_i, \boldsymbol{w})$ or $\max_{\boldsymbol{w}} \prod_{i=1}^{N} p(y_i | \boldsymbol{x}_i, \boldsymbol{w})\, p(\boldsymbol{w})$
- Prediction for $\boldsymbol{x}^*$ is $p(y^* | \boldsymbol{x}^*, \widehat{\boldsymbol{w}})$
- Choose hyper-parameters / models
  - on a held-out validation set
  - by cross-validation
  - on OOB samples (random forest)

Bayesian

- Training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, $\boldsymbol{X}$ for all $\boldsymbol{x}_i$, $\boldsymbol{y}$ for all y_i

$$p(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y}|\boldsymbol{X})} \propto \prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i,\boldsymbol{w})\, p(\boldsymbol{w})$$

- Mean estimate $\mathrm{E}[w]$, uncertainty $\mathrm{Var}(\boldsymbol{w})$ → confidence
- Prediction for $\boldsymbol{x}^*$ is $p(y^*|\boldsymbol{x}^*) = \mathrm{E}_{\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y}}\, p(y^*|\boldsymbol{x}^*,\boldsymbol{w})$
  - by formulas if analytic solution is available
  - by sampling from $p(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y})$ to approximate the prediction
- Choose hyper-parameters / models by comparing $p(\boldsymbol{y}|\boldsymbol{X})$

Frequentist and Bayesian

- Frequentist
  - find a single parameter vector to best fit the training set
  - the best parameters are used to make predictions directly

- Bayesian
  - formulate the full posterior given the training data
  - all the weights are used to make expected predictions
  - where each is scaled by its posterior probability

Bayesian

- Advantages
  - less sensitive to overfitting (expected predictions)
    - particularly with small training sets
  - make use of all the data at once
    - no need to hold out validation data, or repeatedly train and test
    - won't overfit to the held-out set when selecting many parameters
- Disadvantages
  - exact inference is sometimes intractable
  - approximate inference may be inefficient and inaccurate
  - algorithms are sometimes complex

Bayesian formula

$$p(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y}|\boldsymbol{X})}$$

- $p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})$ likelihood
- $p(\boldsymbol{w})$ prior
- $p(\boldsymbol{y}|\boldsymbol{X})$ marginal likelihood or evidence
- $p(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y})$ posterior

- $p(\boldsymbol{y}|\boldsymbol{X}) = \sum_{\boldsymbol{w}} p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})p(\boldsymbol{w})$ or $p(\boldsymbol{y}|\boldsymbol{X}) = \int p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})p(\boldsymbol{w})\, \mathrm{d}\boldsymbol{w}$

Conjugate prior and likelihood

- when $p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})p(\boldsymbol{w})$ has the same form as $p(\boldsymbol{w})$
- simplifies the problem of finding the posterior
  - as needed in Bayesian inference
- allows for exact computation of the evidence
  - $p(\boldsymbol{y}|\boldsymbol{X}) = \sum_{\boldsymbol{w}} p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})p(\boldsymbol{w})$ or $p(\boldsymbol{y}|\boldsymbol{X}) = \int p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})p(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w}$

Suite of useful conjugate priors

	likelihood	conjugate prior
regression	Normal	Normal (for mean)
	Normal	Inverse Gamma (for variance) or Inverse Wishart (covariance)
classification	Binomial	Beta
	Multinomial	Dirichlet
counts	Poisson	Gamma

Bayesian Linear Regression (cont)

- We have two Normal distributions
  - normal likelihood x normal prior
- Their product is also a Normal distribution
  - ***conjugate prior:*** *when product of likelihood x prior results in the same distribution as the prior*
  - *evidence* can be computed easily using the normalising constant of the Normal distribution

$$
\begin{aligned}
p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma^2) &\propto \mathrm{Normal}(\mathbf{w}|\mathbf{0}, \gamma^2 \mathbf{I}_D)\mathrm{Normal}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}_N) \\
&\propto \mathrm{Normal}(\mathbf{w}|\mathbf{w}_N, \mathbf{V}_N)
\end{aligned}
$$

closed form solution for posterior!

Bayesian Linear Regression (cont)

$$p(\mathbf{w}|\mathbf{X},\mathbf{y},\sigma^2) \propto \mathrm{Normal}(\mathbf{w}|\mathbf{0},\gamma^2\mathbf{I}_D)\mathrm{Normal}(\mathbf{y}|\mathbf{X}\mathbf{w},\sigma^2\mathbf{I}_N)$$
$$\propto \mathrm{Normal}(\mathbf{w}|\mathbf{w}_N,\mathbf{V}_N)$$

Note that mean (and mode) are the MAP solution from before

where

$$\mathbf{w}_N = \frac{1}{\sigma^2}\mathbf{V}_N\mathbf{X}'\mathbf{y}$$

$$\mathbf{V}_N = \sigma^2(\mathbf{X}'\mathbf{X} + \frac{\sigma^2}{\gamma^2}\mathbf{I}_D)^{-1}$$

Advanced: verify by expressing product of two Normals, gathering exponents together and ‘completing the square’ to express as squared exponential (i.e., Normal distribution).

Outline

- Review the lecture, background knowledge, etc.
  - MLE, MAP, Bayesian estimates
  - Comparison between Bayesian and frequentist
  - Likelihood, prior, and posterior
  - Conjugate prior and likelihood
    - Bayesian linear regression
  - IPython notebook task: Bayesian linear regression