UNIVERSITY OF SOUTHAMPTON COMP6208W1

SEMESTER 2 EXAMINATION 2020 - 2021

ADVANCED MACHINE LEARNING

DURATION 120 MINS (2 Hours)

This paper contains 4 questions

Answer all questions. Section A (worth 40 marks) is a series of questions with short answers. Section B (worth 60 marks) involves longer questions

An outline marking scheme is shown in brackets to the right of each question.

This examination is worth 80%. The coursework was worth 20%.

This is a take home examination. Detailed instructions are given on line.

University approved calculators MAY be used.

A foreign language dictionary is permitted ONLY IF it is a paper version of a direct 'Word to Word' translation dictionary AND it contains no notes, additions or annotations.

9 page examination paper.

Answers

Copyright 2021 © University of Southampton

Section A

Question A1.

(a) Explain what it means to normalise the input features and why this is often a good thing to do.

Indicative Solution for Question A1(a).

We normalise the features so they all have the same mean and variance. It is often useful as many learning machines will treat larger features as more important. If we don't know which features are important normalising them will remove an unjustified bias.

[5 marks]

(b) What is the purpose of cross-validation and what is its cost?

Indicative Solution for Question A1(b).

Cross-validation allows us to get a more accurate approximation of the generalisation performance while using a large proportion of examples to train the weights. The cost is that we have to train our learning machine multiple times.

[5 marks]

(c) To find the maximum or minimum of a function explain the major ways (i) gradient descent (ii) Newton's method and (iii) quasi-Netwon methods differ in terms of the information they use.

Indicative Solution for Question A1(c).

(i) Gradient descent uses just the gradient $\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - r\,\boldsymbol{\nabla} f(\boldsymbol{x}^{(t)})$

(ii) Newton's method uses both the gradient and the Hessian $\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \mathbf{H}^{-1}\,\boldsymbol{\nabla} f(\boldsymbol{x}^{(t)})$

(iii) Quasi-Newton methods use some method for approximating the Hessian.

[5 marks]

(d) Explain why CNNs capture the structure of typical image datasets.

Indicative Solution for Question A1(d).

CNNs are built from local filters that respond to objects in a translationally invariant way (the same feature maps will be activated by an object irrespective of where it is). Objects tend to be localised.

[5 marks]

Answers

Copyright 2021 © University of Southampton

(e) Describe the Karush-Kuhn-Tucker (KKT) conditions for constrained optimisation.

Indicative Solution for Question A1(e).

The KKT conditions are used when we have inequality constraints. They state that either the Lagrange multiplier is zero and the point found satisfies the constraint and would be an optimum regardless of the constraint, or the Lagrange multiplier is non-zero and the point found lies on the constraint boundary.

[5 marks]

(f) Show that an empirical covariance matrix, $\mathbf{C}$, can be written as $\mathbf{X}\mathbf{X}^\mathsf{T}$ and hence prove that its eigenvalues are non-negative.

Indicative Solution for Question A1(f).

The empirical covariance matrix is given by

$$\mathbf{C} = \frac{1}{m-1}\sum_{i=1}^{m}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})^\mathsf{T}$$

where $\hat{\mu}$ is the empirical mean vector. Defining X as a matrix whose i^{th} column is equal to $(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})/\sqrt{m-1}$ then $\mathbf{C} = \mathbf{X}\mathbf{X}^\mathsf{T}$. If $\boldsymbol{v}$ is an eigenvector of C then $\mathbf{C}\,\boldsymbol{v} = \lambda\,\boldsymbol{v}$. Multiplying on the left by $\boldsymbol{v}^\mathsf{T}$

$$\boldsymbol{v}^\mathsf{T}\mathbf{C}\,\boldsymbol{v} = \boldsymbol{v}^\mathsf{T}\mathbf{X}\mathbf{X}^\mathsf{T}\boldsymbol{v} = \lambda\,\boldsymbol{v}^\mathsf{T}\boldsymbol{v}.$$

Defining $\boldsymbol{u} = \mathbf{X}^\mathsf{T}\boldsymbol{v}$ then

$$\lambda = \frac{\boldsymbol{u}^\mathsf{T}\boldsymbol{u}}{\boldsymbol{v}^\mathsf{T}\boldsymbol{v}} = \frac{\|\boldsymbol{u}\|^2}{\|\boldsymbol{v}\|^2} \geq 0$$

[5 marks]

(g) Prove that $f(x) = \exp(c\,x)$ is a convex-up function.

Indicative Solution for Question A1(g).

To prove this we need to show $f''(x) \geq 0$ for all x. But $f''(x) = c^2\,\exp(c\,x)$. Since $c^2 \geq 0$ and $\exp(c\,x) > 0$ it follows that $f(x)$ is convex-up.

[5 marks]

(h) Explain what the hyper-parameters of a Gaussian process are and why they are relatively easy to learn.

Indicative Solution for Question A1(h).

The hyper-parameters are the mean function $m(\boldsymbol{x})$, the kernel or covariance $k(\boldsymbol{x}, \boldsymbol{y})$ (often the kernel will depend on a scale ℓ that is a hyper-parameter) and the level of noise in the training data, σ. It is relatively straightforward to find good hyper-parameters as we can compute the evidence $\mathbb{P}(\mathcal{D})$ in closed form.

TURN OVER

Copyright 2021 © University of Southampton

Answers

[5 marks]

Answers

Copyright 2021 © University of Southampton

Section B

Question B1.

(a) If $\{X_i | i = 1, 2, \ldots, n\}$ is a set of correlated random variables such that

$$\mathbb{E}[X_i] = \mu \qquad \mathbb{E}[(X_i - \mu)(X_j - \mu)] = \begin{cases} \sigma^2 & \text{if } i = j \\ \rho\,\sigma^2 & \text{if } i \neq j \end{cases}$$

show

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)^2\right] = \rho\,\sigma^2 + \frac{(1-\rho)\,\sigma^2}{n}$$

Indicative Solution for Question B1(a).

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)^2\right] &= \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^{n}(X_i - \mu)\right)^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{i,j=1}^{n}(X_i - \mu)\,(X_j - \mu)\right] = \frac{1}{n^2}\sum_{i,j=1}^{n}\mathbb{E}[(X_i - \mu)\,(X_j - \mu)] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 + \frac{1}{n^2}\sum_{\substack{i,j=1 \\ i\neq j}}^{n}\rho\,\sigma^2 = \frac{1}{n^2}n\,\sigma^2 + \frac{1}{n^2}n\,(n-1)\,\rho\,\sigma^2 \\
&= \rho\,\sigma^2 + \frac{(1-\rho)}{n}\sigma^2
\end{aligned}$$

[10 marks]

(b) Using the result derived in part (a) explain why ensembling many machines can reduce the variance in the bias-variance dilemma if the machine predictions are not heavily correlated. Use this to explain the success of random forest.

Indicative Solution for Question B1(b).

We can think of X_i as the prediction made by a learning machine. For a finite training set this prediction will vary. If our machines are unbiased $\mathbb{E}[X]_i = \mu$, that is, the true value we are trying to predict. The constant ρ is the Pearson correlation between the predictions. By ensembling many machines (high n) we reduce the second term. However, because the machines are correlated our prediction will still vary from the true prediction by $\rho\,\sigma^2$. In random forest we try to make the predictions uncorrelated by basing them on a different set of features and using bootstrapping so they learn on a slightly different training set.

TURN OVER

Copyright 2021 © University of Southampton

Answers

[10 marks]

Answers

Copyright 2021 © University of Southampton

Question B2.

(a) Consider gradient descent

$$x^{(t+1)} = x^{(t)} - r\, f'(x^{(t)})$$

acting on a function $f(x) = x^2$. Explain what happens starting from some point $x^{(0)}$ if (i) $0 < r < \frac{1}{2}$, (ii) $r = \frac{1}{2}$, (iii) $\frac{1}{2} < r < 1$, (iv) $r = 1$ and (v) $r > 1$.

Indicative Solution for Question B2(a).

We note that

$$x^{(t+1)} = (1 - 2\,r)\, x^{(t)}$$

(i) $0 < r < \frac{1}{2}$**:** $x^{(t)}$ **converges monotonically towards the optimum at** $x = 0$**.**

(ii) $r = \frac{1}{2}$**: we jump to the optimum in one step**

(iii) $\frac{1}{2} < r < 1$**:** $x^{(t)}$ **jumps to the other side of the optimum but converges towards it**

(iv) $r = 1$**:** $x^{(t)} = (-1)^t x^{(0)}$

(v) $r > 1$**:** $x^{(t)}$ **diverges from the optimum.**

[10 marks]

(b) Describe the problems that can arise in finding the optimum of a high-dimensional loss function and solution to them.

Indicative Solution for Question B2(b).

Solutions might discuss the following

- **Generally we will not be in a convex landscape so we may have to travel a long way to the optimum. We can follow the gradient.**
- **There will typically be a large range of curvatures meaning there is no clear step size.**
  - **We could use second order methods but in very high-dimensional spaces this is often not practical.**
  - **We can seek to adapt the step size on different variables although this doesn't necessarily help if variables have correlated effects on the loss.**
  - **We can use momentum.**

[10 marks]

Answers

TURN OVER

Copyright 2021 © University of Southampton

Question B3.

(a) Explain how Markov chain Monte-Carlo (MCMC) techniques are used to solve Bayesian inference problems and what problem do they solve.

Indicative Solution for Question B3(a).

MCMC is used to obtain random independent samples from the posterior $f(\boldsymbol{\theta}|\mathcal{D})$. In Bayesian inference

$$f(\boldsymbol{\theta}|\mathcal{D}) = \frac{f(\mathcal{D}|\boldsymbol{\theta})\, f(\boldsymbol{\theta})}{f(\mathcal{D})}.$$

Although we can compute the likelihood, $f(\mathcal{D}|\boldsymbol{\theta})$ and prior $f(\boldsymbol{\theta})$ the posterior marginal likelihood or evidence, $f(\mathcal{D})$, is often intractable. In MCMC we only care about ratios of the joint probability $f(\mathcal{D}, \boldsymbol{\theta}) = f(\mathcal{D}|\boldsymbol{\theta})\, f(\boldsymbol{\theta})$ which we can compute. Usually this joint probability is too complicated to sample using transformation methods and the rejection rate would be far too high to use rejection methods as most parameters have very low likelihoods. MCMC allow us to transition between likely states, although they require many transitions for the samples to be independent.

[10 marks]

(b) Describe the Metropolis algorithm and show that it satisfies detailed balance. Explain why this is important.

Indicative Solution for Question B3(b).

The Metropolis algorithm is an iterative algorithm where we change the parameters θ of our model. We choose a random parameters $\boldsymbol{\theta}' \sim p(\boldsymbol{\theta}'|\boldsymbol{\theta})$ where $p(\boldsymbol{\theta}'|\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{\theta}')$ (we are using Metropolis not Metropolis-Hastings). We update our parameter if

$$\frac{f(\boldsymbol{\theta}', \mathcal{D})}{f(\boldsymbol{\theta}, \mathcal{D})} > U$$

where U is a uniform random variate $U \sim \mathrm{U}(0, 1)$. That is, the transition probability from θ to θ' is

$$W(\boldsymbol{\theta}', \boldsymbol{\theta}) = p(\boldsymbol{\theta}'|\boldsymbol{\theta})\, \min\left(\frac{f(\boldsymbol{\theta}', \mathcal{D})}{f(\boldsymbol{\theta}, \mathcal{D})}, 1\right).$$

We want to converge to the posterior $f(\boldsymbol{\theta}|\mathcal{D})$. This will happen (assuming egodicity) if we satisfy the detailed balance equation

$$W(\boldsymbol{\theta}', \boldsymbol{\theta})\, f(\boldsymbol{\theta}|\mathcal{D}) = W(\boldsymbol{\theta}, \boldsymbol{\theta}')\, f(\boldsymbol{\theta}'|\mathcal{D})$$

or

$$\frac{W(\boldsymbol{\theta}', \boldsymbol{\theta})}{W(\boldsymbol{\theta}, \boldsymbol{\theta}')} = \frac{f(\boldsymbol{\theta}'|\mathcal{D})}{f(\boldsymbol{\theta}|\mathcal{D})} = \frac{f(\boldsymbol{\theta}', \mathcal{D})}{f(\boldsymbol{\theta}, \mathcal{D})}.$$

But by the definition of $W(\boldsymbol{\theta}', \boldsymbol{\theta})$ we see that if $f(\boldsymbol{\theta}', \mathcal{D}) \leq f(\boldsymbol{\theta}, \mathcal{D})$

$$\frac{W(\boldsymbol{\theta}', \boldsymbol{\theta})}{W(\boldsymbol{\theta}, \boldsymbol{\theta}')} = \frac{\frac{p(\boldsymbol{\theta}'|\boldsymbol{\theta})\, f(\boldsymbol{\theta}', \mathcal{D})}{f(\boldsymbol{\theta}, \mathcal{D})}}{p(\boldsymbol{\theta}|\boldsymbol{\theta}') \times 1} = \frac{f(\boldsymbol{\theta}', \mathcal{D})}{f(\boldsymbol{\theta}, \mathcal{D})}$$

Answers

Copyright 2021 © University of Southampton

since by construction $p(\boldsymbol{\theta}'|\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{\theta}')$. We get an identical result if $f(\boldsymbol{\theta}', \mathcal{D}) > f(\boldsymbol{\theta}, \mathcal{D})$. Satisfying detailed balance ensure that eventually we converge to the posterior distribution.

[10 marks]

Answers

END OF PAPER

Copyright 2021 © University of Southampton