70

University of Toronto

STA 414/2104: Midterm , Winter 2015

Name: Jiayu Tang

Student ID: 1000930380.

Read the following instructions carefully:

1. Do not turn the page until told to do so.
2. You may use a single side of 8.5×11 inch aid sheet and a nonprogrammable calculator for this exam.
3. If a question asks you to do some calculations, you must *show your work* to receive full credit.
4. You can use either pen or pencil for the exam. **But please be aware that you are not allowed to dispute any credit after the exam is returned if you use a pencil.**
5. Lastly, enjoy the problems!!!

1. (10 mk) Show that the mode (i.e. the maximum) of the univariate Gaussian distribution

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

is given by μ.

find x, to maximize $N(x|\mu,\sigma^2)$ is to maximize $\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$.

where μ & σ are known

$\Rightarrow$ is equivalent to maximizing $-\frac{1}{2\sigma^2}(x-\mu)^2$. since $h(a) = \exp(a)$ is an increasing function

$\Rightarrow$ that is to minimize $(x-\mu)^2$.

$(x-\mu)^2 \geq 0$, it takes the equality holding only when $x=\mu$.

$\therefore$ When $x=\mu$, $N(x|\mu,\sigma^2)$ is maximized

$\therefore$ Mode of $N(x|\mu,\sigma^2)$ is $x=\mu$.

2. (15 mk) Suppose that we have a dataset of observations: $\mathbf{x} = (x_1, x_2, ..., x_N)^\top$, representing N observations of the scalar binary variable x (so that $x_n \in \{0, 1\}, n = 1, .., N$). Assume that the observations are $i.i.d$ and are drawn from a Bernoulli distribution with parameter μ.

15

(a) (3 mk) Write down the log-likelihood for these N observations.

3

for each observation, $\text{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}$.

Thus for N observations $L(\mu; x_1 \cdots x_n) = \prod_{i=1}^{N} \mu^{x_i} (1-\mu)^{1-x_i}$ since x_i's are i.i.d.

$$\log L(\mu; x_1 \cdots x_n) = \sum_{i=1}^{N} [x_i \log \mu + (1-x_i) \log (1-\mu)].$$

(b) (5 mk) Maximizing the log-likelihood with respect to μ, find the maximum likelihood estimate μ_{ML} of μ (show your derivations).

5

$$\log L = \sum_{i=1}^{N} x_i \log \mu + \sum_{i=1}^{N} (1-x_i) \log (1-\mu)$$

$$= M \log \mu + (N-M) \log (1-\mu).$$ where M is the # of observations that $x_i = 1$, $M = \sum_{i=1}^{N} x_i$

$$\frac{\partial \log L}{\partial \mu} = \frac{M}{\mu} - \frac{N-M}{1-\mu} \overset{\text{set}}{=} 0.$$

get $\mu_{ML} = \frac{M}{N}$. $\frac{M}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$.

Checking the second derivative, $\frac{\partial^2 \log L}{\partial \mu} = -\frac{M}{\mu^2} + \frac{N-M}{(1-\mu)^2} < 0$ at μ_{ML}

$\therefore \mu_{ML} = \frac{M}{N}$ M is defined above.

$= \frac{\sum_{i=1}^{N} x_i}{N}$

(c) (7 mk) Show that the ML estimate of μ is unbiased, i.e. show that:

7

$$\mathbb{E}[\mu_{ML}] = \mu$$

$$E(\mu_{ML}) = E\left(\frac{\sum_{i=1}^{N} x_i}{N}\right) = \frac{1}{N} E\left(\sum_{i=1}^{N} x_i\right) = \frac{1}{N}\left[\sum_{i=1}^{N} E(x_i)\right], \quad E(x_i) = \mu \ \ \forall i > 0$$

$$= \frac{1}{N} \cdot N \cdot \mu = \mu.$$

3. (15 mk) Suppose that we have a dataset of observations: $\mathbf{x} = (x_1, x_2, ..., x_N)^\top$, representing N observations of the scalar variable x. Assume that the observations are *i.i.d* and are drawn from a Gaussian distribution with mean μ and variance σ^2.

Suppose that the variance of a Gaussian is estimated using

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$$

but with maximum likelihood estimate μ_{ML} replaced with a true value μ of the mean. Show that this estimator has the property that its expectation is given by the true variance σ^2, i.e. show that

$$\mathbb{E}[\sigma^2_{ML}] = \sigma^2.$$

$E(\alpha^2_{ML}) = E\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2\right)$ Now.

$= \frac{1}{N}\left[\sum_{i=1}^{N}\{E(x_i - \mu)^2\}\right]$

$E[(x_i - \mu)^2] = E[(x_i - E(x_n))^2]$, by the definition of variance, $E[(x_n - \mu)^2]$ is the true variance α^2 for x, $E[(x_i - \mu)^2] = \alpha^2 \quad \forall i$.

Thus $E(\alpha^2_{ML}) = \frac{1}{N}\sum_{i=1}^{N}\alpha^2 = \alpha^2$.

15

Remark

4. (20 mk) Consider a regression model $y(\mathbf{x})$ with a squared loss function. In this case the expected loss can be written as:

12

$$\mathbb{E}[L] = \int \int (y(\mathbf{x}) - t)^2 p(\mathbf{x}, t) d\mathbf{x} dt$$

Show that the function $y(\mathbf{x})$ for which this expected loss is minimized is given by the conditional expectation $\mathbb{E}[t|\mathbf{x}] = \int t p(t|\mathbf{x}) dt$.

$$E[L] = \iint y(x)^2 \cdot p(x,t)\, dx\, dt + \iint t^2 p(x,t)\, dx\, dt - 2\iint y(x)\, t\, p(x,t)\, dx\, dt$$

$$= \int y(x)^2 \cdot p(x)\, dx + \int t^2 \cdot p(t)\, dt - 2 \int y(x)\, dx \int t\, p(t|x) \cdot p(x) \cdot dt$$

$$= \int y(x)^2 \cdot p(x)\, dx + E[t^2] - 2 \int y(x) p(x) \cdot E[t|x] \cdot dx.$$

$$= E_t(t^2) + E_x[y(x)^2 - 2 y(x) E(t|x)]$$

20/20

∴ To minimize $E[L]$ is to minimize $E_x[y^2(x) - 2E(t|x) \cdot y(x)]$

$\forall x$, $y(x) - 2E[t|x] \cdot y(x)$ is minimized at $y(x) = E(t|x)$

Thus $E_x[y(x) - 2E(t|x) \cdot y(x)] = \int [y(x) - 2y(x)E(t|x)]\, p(x)\, dx$ is always minimized at $y(x) = E(t|x)$.

∴ $E[L]$ is minimized at $y(x) = \int t p(t|x)\, dt = E[t|x]$.

Based on this, $y(x) = 0$ also works. ← ~~correct~~ ✓

5. (20 mk) Consider a linear basis function regression model. We observe a dataset of inputs 15 $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ and corresponding real-valued target values $\mathbf{t} = \{t_1, ..., t_N\}$. Assuming that the observations are *i.i.d*, the likelihood function can be written as:

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^\top \phi(\mathbf{x}_n), \sigma^2)$$

Consider a zero mean Gaussian prior governed by a single precision parameter α:

evidence approximation. $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}I)$

(a) (3 mk) Write down the form of the marginal likelihood for this model.

1 $p(t|X, w, \alpha^2) = \int \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\alpha^2}} \exp\left(-\frac{1}{2\alpha^2}(X_n - w^T\phi(X_n)^2\right) p(w|\alpha)\, dw.$

$= (2\pi\alpha^2)^{-\frac{N}{2}} \cdot \exp\left\{-\frac{1}{2\alpha^2}\sum_{n=1}^{N}(X_n + w^T\phi(X_n)^2\right\}.$

$p(t|X, \alpha, \beta) = \int p(t|X, w, \beta)\, p(w|\alpha)\, dw.$

(b) (5 mk) Write down the form of predictive distribution of t for a new input $\mathbf{x}^*$.

2 Set $\beta^{-1} = \alpha^2$, $p(t^*|t, X^*, X, \alpha, \beta) = \underbrace{N(t^*|m_N^T\phi(X^*), \alpha^2_N(X^*)}$

Bayesian Approach. where $\alpha^2_N(X^*) = \frac{1}{\beta} + \phi(X^*)^T S_N \phi(X^*)$. ↳ Indicate where this comes from.

$S_N^{-1} = \alpha I + \beta\Phi^T\Phi \quad m_N = \beta S_N \Phi^T t$

$\Phi = \begin{pmatrix} \phi_0(X_1) & \phi_1(X_1) & \cdots & \phi_{M-1}(X_1) \\ \phi_0(X_2) & \phi_1(X_2) & \cdots & \phi_{M-1}(X_2) \\ \vdots & & & \\ \phi_0(X_N) & \cdots & & \phi_{M-1}(X_N) \end{pmatrix}$

(c) (12 mk) Show that maximizing the log of the posterior distribution with respect to $\mathbf{w}$ is equivalent to the minimization of the sum-of-squares error function with the addition of a quadratic regularization term, with regularization parameter $\lambda = \alpha * \sigma^2$.

12

the posterior fn $p(w|D)$

since $p(w|D) \propto p(w) \cdot p(D|w)$.

$p(w|D) = C \cdot \exp\left[-\frac{\beta}{2}\sum_{n=1}^{N}(t_n - w^T\phi(x_n))^2\right] \cdot \exp\left[-\frac{\alpha}{2}w^Tw\right]$. where C is a constant

$\log p(w|D) = -\frac{\beta}{2}\sum_{n=1}^{N}(t_n - w^T\phi(x_n))^2 - \frac{\alpha}{2}w^Tw + \text{constant}$

To maximize $\log p(w|D)$ w.r.t w is to minimize

$$-\frac{\beta}{2}\sum_{n=1}^{N}(t_n - w^T\phi(x_n))^2 - \frac{\alpha}{2}w^Tw$$

is to minimize

$$\beta\left[\frac{1}{2}\sum_{n=1}^{N}(t_n - w^T\phi(x_n))^2 + \frac{1}{2}\frac{\alpha}{\beta}w^Tw\right], \quad \beta^{-1} = \sigma^2 > 0.$$

Thus it is equivalent to minimize the error function.

$\frac{1}{2}\sum_{n=1}^{N}(t_n - w^T\phi(x_n))^2$ is the sum of the square error for the model

plus the regularization term.

$\frac{1}{2}\lambda w^Tw$ where $\lambda = \frac{\alpha}{\beta} = \alpha\sigma^2$.

6. (20 mk) Consider a Bayesian logistic regression model for a binary classification problem, parametrized by $\mathbf{w}$. Consider a dataset $\{\mathbf{x}_n, t_n\}$, where $t_n \in \{0, 1\}$ and $n = 1, ..., N$. Let us consider a zero mean Gaussian prior over $\mathbf{w}$ governed by a single precision parameter α:

20

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}I) \tag{1}$$

(a) (5 mk) Write down the posterior distribution over $\mathbf{w}$ for the Bayesian logistic regression model. (Hint: posterior is proportional to the likelihood function times prior)

5

$p(w|x,t) \propto p(t|x,w) \cdot p(w|\alpha)$. $\quad p(w|\alpha) = N(w|0, \alpha^{-1}I)$.

Then $\log p(w|x,t) = \log p(t|x,w) + \log p(w|\alpha) + \text{Constant}$

$$= -\frac{1}{2} w^T \alpha I w + \sum_{n=1}^{N} [t_n \ln y_n + (1-t_n)\ln(1-y_n)] + \text{constant}.$$

Thus $p(w|x,t) \propto \exp\{-\frac{1}{2}\alpha w^T w\} \cdot \prod_{i=1}^{N} y_n^{t_n} (1-y_n)^{(1-t_n)}$

where $y_n = \sigma(w^T x_n) = \frac{1}{1+\exp(-w^T x_n)}$

The corresponding Gaussian approximation is. Not needed.

$q(w) = N(w|w_{map}, S_N^{-1})$ $\quad S_N^{-1} = \alpha I + \sum_n y_n(1-y_n) x_n x_n^T$ $\quad w_{map}$ is the mode of the posterior

(b) (5 mk) Show that maximizing the log of the posterior distribution with respect to $\mathbf{w}$ is equivalent to the minimization of the cross-entropy error function function with the addition of a quadratic regularization term (this corresponds to regularized cross-entropy error function).

5

To maximiz $\log p(w|x,t) = -\frac{\alpha}{2} w^T w + \sum_{n=1}^{N} \{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} + \text{constant}$

is to minimize $-\sum_{n=1}^{N} \{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} + \frac{\alpha}{2} w^T w.$

$$= \sum_{n=1}^{N} E_n + \frac{\alpha}{2} w^T w$$

where $E_n = -[t_n \ln y_n + (1-t_n)\ln(1-y_n)]$

$$E_n = E_n(y_n) = \begin{cases} -\ln y_n & \text{if } t_n = 1 \\ -\ln(1-y_n) & \text{if } t_n = 0. \end{cases}$$

$\sum E_n$ is the cross-entropy error in total.

and $\frac{\alpha}{2} w^T w$ is the corresponding regularization term with $\lambda = \alpha$.

(c) (10 mk) Compute the derivative of the regularized cross-entropy function with respect to parameter vector **w** (which is needed for parameter estimation).

10

$$E[L] = -\sum_{n=1}^{N} \{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} + \frac{\alpha}{2} w^T w.$$

$$= \sum_{n=1}^{N} E_n + \frac{\alpha}{2} w^T w \quad \checkmark$$

$$\frac{dE_n}{dy_n} = \frac{y_n - t_n}{y_n(1-y_n)} \qquad \frac{dy_n}{dw} = y_n(1-y_n)\cdot x_n \quad \text{since} \quad \frac{dy_n}{dw} = \frac{d\sigma(w^T x_n)}{dw}$$

$$= y_n(1-y_n)x_n.$$

$$\therefore \frac{dE_n}{dw} = \frac{dE_n}{dy_n}\cdot\frac{dy_n}{dw} = (y_n - t_n)\cdot x_n.$$

$$\nabla_w [w^T w] = 2\cdot w. \quad \checkmark$$

$$\text{Thus } \nabla_w E[L] = \sum_{n=1}^{N}(y_n - t_n)\cdot x_n + \alpha\cdot w.$$

95/100