Homework #8

TA in charge: Yao-Nan Chen

RELEASE DATE: 12/27/2010

DUE DATE: 01/03/2011, 4:00 pm IN CLASS

TA SESSION: 12/30/2010, 6:00 pm IN R110

The homework is OPTIONAL. That is, if you choose to turn it in, your homework score would be calculated over HW1 to HW8; otherwise your homework score would be calculated over HW1 to HW7. In both cases, we will use the equation

$$\frac{\textbf{your best homework} * 1.5 + \textbf{your worse homework} * 0.5 + \sum(\textbf{your other homework})}{\textbf{\# of homework}}$$

Please make a choice BEFORE the TA grades HW8. If you choose to not turn in HW8, we still encourage you to discuss the solutions with your classmates or TAs.

Unless granted by the instructor in advance, you must turn in a hard copy of your solutions (without the source code) for all problems. For problems marked with (*), please follow the guidelines on the course website and upload your source code to designated places.

Any form of cheating, lying, or plagiarism will not be tolerated. Students can get zero scores and/or fail the class and/or be kicked out of school and/or receive other punishments for those kinds of misconducts.

Discussions on course materials and homework solutions are encouraged. But you should write the final solutions alone and understand them fully. Books, notes, and Internet resources can be consulted, but not copied from.

Since everyone needs to write the final solutions alone, there is absolutely no need to lend your homework solutions and/or source codes to your classmates at any time. In order to maximize the level of fairness in this class, lending and borrowing homework solutions are both regarded as dishonest behaviors and will be punished according to the honesty policy.

You should write your solutions in English with the common math notations introduced in class or in the problems. We do not accept solutions written in any other languages.

8.1 A Bayesian View of Logistic Regression

(1) (10%) Prove that logistic regression (see Problem 4.6) equivalently minimizes the following error function:

$$E_{ce}(\mathbf{w}) = -\sum_{n=1}^{N} \left(\frac{1+y_n}{2} \ln \frac{1+\tanh(\frac{1}{2}\mathbf{w} \bullet \mathbf{x}_n)}{2} + \frac{1-y_n}{2} \ln \frac{1-\tanh(\frac{1}{2}\mathbf{w} \bullet \mathbf{x}_n)}{2} \right).$$

The error function is usually called the *cross-entropy*.

(2) (10%) Assume that the universe generates an example $(\mathbf{x}, y)$ by the following procedure:

(a) generate $\mathbf{x}$ from some probability density function $P(\mathbf{x})$

(b) use some fixed $\mathbf{w}^u$ (including $w_0^u = 1$) to evaluate $\rho = \mathbf{w}^u \bullet \mathbf{x}$

(c) evaluate $Q_+ = \exp\left(\frac{\rho}{2}\right)$ and $Q_- = \exp\left(-\frac{\rho}{2}\right)$

(d) generate $y \in \{+, -\}$ with the probability distribution $Q_y/(Q_+ + Q_-)$

If each $(\mathbf{x}_n, y_n)$ within $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ is generated i.i.d from the procedure above, what is the likelihood $P\big(\mathcal{D} | \mathbf{w} = \mathbf{w}^u\big)$?

(3) (10%) Prove that logistic regression equivalently gives the maximum likelihood estimate of $\mathbf{w}^u$.

(4) (10%) If we take the solution $\hat{\mathbf{w}}$ from logistic regression as our estimate of the underlying $\mathbf{w}^u$. Show that $\frac{Q_+}{Q_+ + Q_-}$ can be estimated by $\frac{1}{1+\exp(-\hat{\mathbf{w}} \bullet \mathbf{x})}$.

Note that the cross-entropy error function is often used to design other learning algorithms (for example, some Neural Networks). In addition, the logistic function in (4) is often used to obtain "probability estimates" of linear classifiers.

8.2 A Bayesian View of Linear Regression

(1) (10%) Assume that the universe generates an example $(\mathbf{x}, y)$ by the following procedure:

(a) generate $\mathbf{x}$ from some probability density function $P(\mathbf{x})$

(b) use some fixed $\mathbf{w}^u$ (including $w_0^u = 1$) to evaluate $\rho = \mathbf{w}^u \bullet \mathbf{x}$

(c) generate $y \in \mathbb{R}$ from ρ by the probability density function $P(y|\,\rho) = \frac{1}{\sqrt{2\pi}} \exp\left(-(y-\rho)^2\right)$

If each $(\mathbf{x}_n, y_n)$ within $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ is generated i.i.d from the procedure above, what is the likelihood $P\big(\mathcal{D}|\,\mathbf{w} = \mathbf{w}^u\big)$?

(2) (10%) Prove that least-square linear regression (see Problem 4.5) equivalently gives the maximum likelihood estimate of $\mathbf{w}^u$.

(3) (10%) Assume that the universe generates an example $(\mathbf{x}, y)$ by the following procedure:

(a) generate some $\mathbf{w}^u$ from

$$P(\mathbf{w}^u) = \frac{1}{(\sqrt{2\pi})^{d+1} \cdot \sigma^{d+1}} \cdot \exp\left(-\frac{\|\mathbf{w}^u\|^2}{2\sigma^2}\right)$$

(b) generate $\mathbf{x}$ from some probability density function $P(\mathbf{x})$

(c) use the $\mathbf{w}^u$ to evaluate $\rho = \mathbf{w}^u \bullet \mathbf{x}$

(d) generate $y \in \mathbb{R}$ from ρ by the probability density function $P(y|\,\rho) = \frac{1}{\sqrt{2\pi}} \exp\left(-(y-\rho)^2\right)$

If each $(\mathbf{x}_n, y_n)$ within $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ is generated i.i.d from the procedure above, what is the posterior probability $P\big(\mathbf{w} = \mathbf{w}^u|\,\mathcal{D}\big)$?

(4) (10%) Prove that regularized linear regression (see Problem 5.3) equivalently gives the maximum posterior estimate of $\mathbf{w}^u$. What is the relationship between λ (in Problem 5.3) and σ (here)?

(5) (10%) Prove or disprove that on any new input vector $\mathbf{x}^{test}$, the prediction (see Problem 5.3)

$$g(\mathbf{x}^{test}) = \mathbf{w}_{reg}(\lambda) \bullet \mathbf{x}^{test}$$

happens to correspond to the Bayes estimate (in the regression sense) with respect to the posterior probability derived above.

(6) (10%) Assume that you have a strong *prior* belief that the universe generates an example $(\mathbf{x}, y)$ by the following procedure:

(a) generate some $\mathbf{w}^u$ that is similar to a "prior" target $\hat{\mathbf{w}}$ by

$$P(\mathbf{w}^u) = \frac{1}{(\sqrt{2\pi})^{d+1} \cdot \sigma^{d+1}} \cdot \exp\left(-\frac{\|\mathbf{w}^u - \hat{\mathbf{w}}\|^2}{2\sigma^2}\right)$$

(b) generate $\mathbf{x}$ from some probability density function $P(\mathbf{x})$

(c) use the $\mathbf{w}^u$ to evaluate $\rho = \mathbf{w}^u \bullet \mathbf{x}$

(d) generate $y \in \mathbb{R}$ from ρ by the probability density function $P(y|\,\rho) = \frac{1}{\sqrt{2\pi}} \exp\left(-(y-\rho)^2\right)$

If each $(\mathbf{x}_n, y_n)$ within $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ is generated i.i.d from the procedure above, what is the posterior probability $P\big(\mathbf{w} = \mathbf{w}^u|\,\mathcal{D}\big)$?

(7) (Bonus 5%) Derive a closed-form solution that gives the maximum posterior estimate of $\mathbf{w}^u$ in Problem 8.2(6).