Exam Pattern Recognition
Wednesday, April 17, 2019
13.30-16.30 hours
Answer Indications

Question 1: General Principles (16 points)

(a) QUAD. QUAD encompasses LIN (setting $w_2 = 0$ in QUAD gives LIN) and has one extra parameter to fit to the data. Hence QUAD will always fit the training data at least as well as LIN, and almost always better. In other words: even though the true function is linear, the quadratic model still fits better to the training data, because it can adapt better to the noise.

(b) LIN. The true function is linear, so both LIN and QUAD are unbiased, but LIN has the smaller variance. The downside of QUAD being able to adapt itself better to the training data is increased variance.

(c) QUAD. See answer to (a) and now the true function is quadratic, so LIN is biased as well.

(d) Mean Squared Error = squared bias + variance. The trade-off is between (squared) bias and variance (see section 3.2 of the book of Bishop). LIN is biased, but has the smaller variance. Since the sample size is fairly small ($N = 11$) the variance component could be substantial. Even though the true model is quadratic, it could still be better to use the linear model in this case.

Question 2: Logistic Regression (20 points)

(a) The probability of acceptance is:

$$P(t = 0) = \frac{1}{1 + e^{-4.1264+5.3714\times 0.33}} = 0.9132388$$

(b) The sign of the coefficient is positive. This makes sense. The higher debt divided by income, the more likely it is that the person will not be able to pay his or her debt. Since the bank obviously wants to avoid this risk, a higher debt to income ratio makes it more likely that the application will be denied.

(c) For white applicants, deny the application if:

$$\begin{aligned} -4.1264 + 5.3714x &> 0 \\ 5.3714x &> 4.1264 \\ x &> 0.7682169 \end{aligned}$$

For black applicants, deny the application if:

$$\begin{aligned} -4.1264 + 5.3714x + 1.2733 &> 0 \\ 5.3714x &> 4.1264 - 1.2733 \\ x &> 0.5311651 \end{aligned}$$

Here x denotes the debt income ratio.

(d) Supporting claim: white applicants are allowed to have a higher debt income ratio before they get denied than black applicants. We can also see this by looking at the positive sign of the coefficient of `black applicant?`. Since this sign is positive, a black person with the same debt to income ratio as a white person, has a higher probability to be rejected. Since the coefficient estimate is significant ($p < 2 \times 10^{-16}$) we can claim that this effect is real, and not due to chance. Against: there could be other relevant criteria (not included in the model) on which black applicants tend to score worse than white applicants. Hence, the analysis is not conclusive.

Question 3: Support Vector Machines (24 points)

(a) Pick any support vector $\mathbf{x}_s$, and use the given formula for computing b.
We pick support vector $\mathbf{x}_4$:

$$b = -1 + \frac{1}{4}[4 \quad 4]\begin{bmatrix} 4 \\ 4 \end{bmatrix} - \frac{1}{8}[4 \quad 4]\begin{bmatrix} 7 \\ 5 \end{bmatrix} - \frac{1}{8}[4 \quad 4]\begin{bmatrix} 5 \\ 7 \end{bmatrix} = -5$$

(b) We compute:

$$\mathbf{w} = \sum_{n=1}^{N} a_n t_n \mathbf{x}_n = -\frac{1}{4}\begin{bmatrix} 4 \\ 4 \end{bmatrix} + \frac{1}{8}\begin{bmatrix} 5 \\ 7 \end{bmatrix} + \frac{1}{8}\begin{bmatrix} 7 \\ 5 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$$

Hence, the decision boundary is:

$$y(x_1, x_2) = -5 + \tfrac{1}{2}x_1 + \tfrac{1}{2}x_2 = 0$$

(c) Fill in $x_1 = 7$ and $x_2 = 4$:

$$y(7, 4) = -5 + \frac{1}{2} \times 7 + \frac{1}{2} \times 4 = \frac{1}{2}$$

Since y is positive, we predict class $+1$.

(d) The decision boundary is given by:

$$y(x_1, x_2) = x_1 + x_2 - 8 = 0$$

(e) First we scale the coefficients so that $t_i y(\mathbf{x}_i) = 1$ for the points closest to the decision boundary. The point (5,7) is one of the points on the margin boundary, and we will use it to scale the coefficients. With the current coefficients we have:

$$y(5, 7) = 5 + 7 - 8 = 4,$$

so we have to divide the coefficients by 4. The rescaled decision boundary becomes:

$$y(x_1, x_2) = \tfrac{1}{4}x_1 + \tfrac{1}{4}x_2 - 2 = 0$$

The value of the slack variable for $\mathbf{x}_4$ is:

$$y(4, 4) = \tfrac{1}{4} \times 4 + \tfrac{1}{4} \times 4 - 2 = 0 \Rightarrow \xi_4 = |-1 - 0| = 1$$

(f) For (b) we get:

$$\frac{1}{2}\mathbf{w}^\top\mathbf{w} + C\sum_{n=1}^{N}\xi_n = \tfrac{1}{2}[\tfrac{1}{2} \quad \tfrac{1}{2}]\begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} + \tfrac{1}{4} \times 0 = \tfrac{1}{4}$$

For (d) we get:

$$\frac{1}{2}\mathbf{w}^\top\mathbf{w} + C\sum_{n=1}^{N}\xi_n = \tfrac{1}{2}[\tfrac{1}{4} \quad \tfrac{1}{4}]\begin{bmatrix} \frac{1}{4} \\ \frac{1}{4} \end{bmatrix} + \tfrac{1}{4} \times 1 = \tfrac{1}{16} + \tfrac{1}{4}$$

So the solution at (b) has the smaller loss.