COMPGI20
Introduction to Supervised Learning
Solutions to Analytical Exercises

March 6, 2017

1 Introduction, Probability

1.1 Qualitative Understanding

1.1.1 Question 1

The task is regression analysis.

1.1.2 Question 2

Indicative useful inputs are:

- Period in the year - can be easily measured
- Time of day - can be easily measured
- Weather conditions - data for this can be obtained from the meteorological office
- Passenger traffic during previous hours per line - can be obtained from number of validated tickets.

1.2 Probability

1.2.1 Question 1

- Box r - 30 apples, 4 oranges, 3 limes. Total 37.
- Box b - 1 apples, 1 oranges, 0 limes. Total 2.
- Box g - 3 apples, 3 oranges, 4 limes. Total 10.

The prior probabilities are given by:

$$p(r) = 0.1$$
$$p(b) = 0.3$$
$$p(g) = 0.6$$

Therefore, the probability that an apple is selected is:

$$\begin{aligned} p(a) &= p(a|r)p(r) + p(a|b)p(b) + p(a|g)p(g) \\ &= \frac{30}{37}(0.1) + \frac{1}{2}(0.3) + \frac{3}{10}(0.6) \\ &= 0.411 \end{aligned}$$

1.2.2 Question 2

By applying Bayes' rule:

$$\begin{aligned} p(r|o) &= \frac{p(o|r)p(r)}{p(o)} \\ &= \frac{\frac{4}{37}(0.1)}{p(o|r)p(r) + p(o|b)p(b) + p(o|g)p(g)} \\ &= \frac{\frac{4}{37}(0.1)}{\frac{4}{37}(0.1) + \frac{1}{2}(0.3) + \frac{3}{10}(0.6)} \\ &= 0.0317 \end{aligned}$$

1.3 Probability

1.3.1 Question 1

$$
\begin{aligned}
E(X) &= \left(\frac{1+2+3+4+5+6}{6}\right)(p) + 6(1-p) \\
&= \frac{21}{6}(p) + 6 - 6p \\
&= 6 - \frac{5}{2}p
\end{aligned}
$$

1.3.2 Question 2

$$
Var(X) = E(X^2) - [E(X)]^2
$$

Where,

$$
\begin{aligned}
E(X^2) &= \left(\frac{1^2+2^2+3^2+4^2+5^2+6^2}{6}\right)(p) + 6^2(1-p) \\
&= \left(\frac{1+4+9+16+25+36}{6}\right)(p) + 36(1-p) \\
&= \frac{91}{6}(p) + 36 - 36p \\
&= 36 - \left(\frac{216-91}{6}\right)(p) \\
&= 36 - \frac{125}{6}p
\end{aligned}
$$

And,

$$
\begin{aligned}
[E(X)]^2 &= \left(6 - \frac{5}{2}p\right)^2 \\
&= 36 - 30p - 6.25p^2
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
Var(X) &= \left(36 - \frac{125}{6}p\right) - \left(36 - 30p - 6.25p^2\right) \\
&= \left(\frac{180-125}{6}\right)(p) - 6.25p^2 \\
&= \frac{55}{6}p - 6.25p^2
\end{aligned}
$$

2 Linear Regression

2.1 Qualitative Understanding

2.1.1 Question 1

Recall that $\mathbf{x} = (y, m, c, a)$ and consider the two cases:
Case 1: $\mathbf{w} = (-1, -2, 1, -10)$
Case 2: $\mathbf{w} = (-1, -2, 10, -1)$

Case 1 implies that for a unit change in condition, the price will change by £1000 and will decrease by £10000 if the car has been involved in an accident.
Case 2 implies that for a unit change in condition, the price will change by £10000 and will decrease by £1000 if the car has been involved in an accident.

Therefore, intuitively, Case 1 is the more reasonable option as from experience, whether or not the car has been involved in an accident has a much greater multiplier effect on price than a slight difference in condition.

2.1.2 Question 2

If only 3 previous sales records are available, 3 equations can be written for 4 unknowns. Therefore, the system is under-determined and cannot be accurately solved. As such, the regression results are not trustworthy.

2.1.3 Question 3

Zero training error, is a sign of over-fitting. This is therefore not a good result as we have too many solutions for the same problem. The approach can be improved by penalizing complexity through regularization and then cross-validating to optimize our complexity-penalization hyper-parameter (λ).

2.2 Least Squares Estimation

Done in class-room.

2.3 Regularization and Priors

2.3.1 Question 1

From Bayes' theorem,

$$p(\mathbf{w}|S) = \frac{p(S|\mathbf{w})p(\mathbf{w})}{p(S)}$$

Where $p(S)$ is the marginal probability given by:

$$p(S) = \sum_i p(S|w_i)p(w_i)$$

Since $p(S)$ does not depend on the value of $\mathbf{w}$ as this is summed over, the value of $\mathbf{w}$ which maximizes $p(\mathbf{w}|S)$ will be equal to the value of $\mathbf{w}$ which maximizes $p(S|\mathbf{w})p(\mathbf{w})$
Therefore,

$$\mathbf{w}_{MAP} = argmax_w p(\mathbf{w}|S) = argmax_w p(S|\mathbf{w})p(\mathbf{w})$$

2.3.2 Question 2

$$\begin{aligned}
\mathbf{w}_{MAP} &= argmax_w p(\mathbf{w}|S) \\
&= argmax_w log(p(\mathbf{w}|S)) \\
&= argmax_w log(pS|\mathbf{w})p(\mathbf{w})) \\
&= argmax_w (log(p(\mathbf{w}) + log(p(\mathbf{S}|w)) \\
&= argmax_w \left(-\frac{\mathbf{S}}{2}log(2\pi) - frac\lambda 2\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{N} \left(-frac12log(2\pi\sigma^2) - frac(y_n\mathbf{w}^T\mathbf{x}_i)^2 2\sigma^2 \right) \right)
\end{aligned}$$

Therefore converting $argmax$ to $argmin$ and ignoring constants:

$$\mathbf{w}_{MAP} = argmax_w \left(\frac{\lambda}{2}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{N} (f_w(x_i) - y_i)^2 \right)$$

2.3.3 Question 3

For the Maximum A Posteriori, $a = \sigma^2$ and $a = \lambda^{-1}$ which implies that $\lambda = \frac{1}{\sigma^2}$.

2.4 Linear Discriminants for multiple classes

2.4.1 Question 1

The equation of the hyperplane separating class j and k is defined as the line at which the plane of intersection between f_k and f_j is equal to zero as this is the boundary at which one hyperplane switches from going to smaller than greater the other . That is:

$$(\mathbf{w}_k^T - \mathbf{w}_j^T)\mathbf{x} + (b_k - b_j) = 0$$

2.4.2 Question 2

Given,

$$\begin{aligned} f_j(\mathbf{x}_A) &= \mathbf{w}_j^T \mathbf{x}_A + \mathbf{b}_j > f_k(\mathbf{x}_A) \\ f_j(\mathbf{x}_B) &= \mathbf{w}_j^T \mathbf{x}_B + \mathbf{b}_j > f_k(\mathbf{x}_B) \end{aligned}$$

Then solving,

$$\begin{aligned} f_j(\lambda \mathbf{x}_A + (1-\lambda)\mathbf{x}_B) &= \mathbf{w}_k^T(\lambda \mathbf{x}_A + (1-\lambda)\mathbf{x}_B) + \mathbf{b}_K \\ &= \lambda \mathbf{w}_k^T \mathbf{x}_A + \mathbf{w}_k^T \mathbf{x}_B - \lambda \mathbf{w}_k^T \mathbf{x}_B + \mathbf{b}_k \\ &= f_k(\mathbf{x}_B) - \lambda f_k(\mathbf{x}_B) + \lambda f_k(\mathbf{x}_A) \\ &= \lambda f_k(\mathbf{x}_A) + (1-\lambda) f_k(\mathbf{x}_B) \end{aligned}$$

Since $0 \leq \lambda \leq 1$, this implies that:

$$\begin{aligned} f_k(x_A) > f_j(x_A) \quad \forall k \neq j \\ f_k(x_B) > f_j(x_B) \quad \forall k \neq j \end{aligned}$$

3 Logistic Regression

3.1 Optimization for Logistic Regression

3.1.1 Question 1

$$H(\mathbf{w}) = \mathbf{X}^T \mathbf{R} \mathbf{X}$$

Where,

$$\begin{aligned} \mathbf{X} &= [x_1, x_2, ..., x_N]^T \\ R_{i,i} &= g(\mathbf{w}^T \mathbf{x}_i)(1 - g(\mathbf{w}^T \mathbf{x}_i)) \end{aligned}$$

Solving:

$$\begin{aligned} \mathbf{X}^T \mathbf{R} &= [x_1, x_2, ..., x_N] \begin{bmatrix} R_{11} & 0 & \dots & 0 \\ 0 & R_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & R_{N,N} \end{bmatrix} \\ &= \begin{bmatrix} x_1 R_{11} & 0 & \dots & 0 \\ 0 & x_2 R_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & x_N R_{N,N} \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \mathbf{X}^T \mathbf{R} \mathbf{X} &= \begin{bmatrix} x_1 R_{11} & 0 & \dots & 0 \\ 0 & x_2 R_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & x_N R_{N,N} \end{bmatrix} [x_1, x_2, ..., x_N]^T \\ &= \begin{bmatrix} x_1 R_{11} x_1 & x_2 R_{22} x_2 & \dots & x_N R_{N,N} x_N \end{bmatrix}^T \\ &= \sum_{i=1}^{N} x_i R_{i,i} x_i \\ &= \sum_{i=1}^{N} x_i g(\mathbf{w}^T \mathbf{x}_i)(1 - g(\mathbf{w}^T \mathbf{x}_i)) x_i \end{aligned}$$

3.1.2 Question 2

The gradient of the loss function would change by adding the l_2 regularization term as it would now need to consider the first derivative of this term such that it would now contain the term:

$$2\lambda \mathbf{w}$$

Similarly, the Hessian matrix would contain the second derivative of the regularization term such that:

$$H(\mathbf{w}) = \mathbf{X}^T\mathbf{R}\mathbf{X} + 2\lambda\mathbf{I}$$

3.2 Multi-class Classification and Logistic Regression

Given $C = 2$, therefore $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2]$ which gives:

$$\begin{aligned} P(y = 1|\mathbf{x}, \mathbf{W}) &= \frac{exp(\mathbf{w}_1^T\mathbf{x})}{exp(\mathbf{w}_1^T\mathbf{x}) + exp(\mathbf{w}_2^T\mathbf{x})} \\ &= \frac{1}{1 + exp((\mathbf{w}_2^T - \mathbf{w}_1^T)\mathbf{x})} \end{aligned}$$

Comparing exponents gives:

$$\begin{aligned} -\mathbf{w}^T\mathbf{x} &= (\mathbf{w}_2^T - \mathbf{w}_1^T)\mathbf{x} \\ -\mathbf{w}^T &= \mathbf{w}_2^T - \mathbf{w}_1^T \\ \mathbf{w}^T &= \mathbf{w}_1^T - \mathbf{w}_2^T \end{aligned}$$

which is valid since $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2]$

4 SVMs

4.1 Geometry

4.1.1 Question 1

Given $f(x) = \mathbf{w}^T\mathbf{x} + b$ with a decision boundary defined by $\mathbf{d} = \mathbf{w}^T\mathbf{x} + b = 0$, the direction of the weight vector $\mathbf{w}$ is perpendicular to the decision boundary where $< \mathbf{w}, \mathbf{d} >= 0$.
Taking two arbitrary points $\mathbf{x}_i, \mathbf{x}_j$, we can infer that:

$$\begin{aligned} \mathbf{w}^T\mathbf{x}_i + b = 0 \\ \mathbf{w}^T\mathbf{x}_j + b = 0 \end{aligned}$$

Solving for b:

$$\begin{aligned} \mathbf{w}^T\mathbf{x}_i + b &= \mathbf{w}^T\mathbf{x}_j + b \\ \mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j &= 0 \\ \mathbf{w}^T(\mathbf{x}_i - \mathbf{x}_j) &= 0 \end{aligned}$$

Since $(\mathbf{x}_i - \mathbf{x}_j)$ represents the direction vector of the decision boundary and we have proved that the inner product $< \mathbf{w}, (\mathbf{x}_i - \mathbf{x}_j) >= 0$, then it follows that the direction of the weight vector is perpendicular to the decision boundary.

4.2 Representer Theorem

Given a set of feature-label pairs, $\{\mathbf{x}_i, y_i\}$, the training objective for SVMs can be expressed as follows:

$$C(\mathbf{w}) = \sum_{i=1}^{N} \max(0, 1 - y_i\mathbf{w}^T\mathbf{x}_i) + \lambda\mathbf{w}^T\mathbf{w}.$$

This criterion is composed of two terms: the empirical loss, $\sum_{i=1}^{N} \max(0, 1 - y_i\mathbf{w}^T\mathbf{x}_i)$ and the regularizer, $\mathbf{w}^T\mathbf{w}$.
We will prove the representer theorem by reduction to the absurd.

We denote by $\mathbf{w}_{\mathcal{X}}$ the weight vector that minimizes Eq. **??**, while lying on the span of the training points $\mathcal{X} = \text{span}(\{\mathbf{x}_i\},\ i \in 1, \ldots, N)$:

$$\mathbf{w}_{\mathcal{X}} = \sum_{i=1}^{N} a_i \mathbf{x}_i$$

Let us assume that there exists a better weight vector, $\mathbf{w}_B$ that yields a lower score of **??**. By the definition of $\mathbf{w}_{\mathcal{X}}$ this could only happen if $\mathbf{w}_B$ contains a component that is outside the span of $\mathcal{X}$.

Decomposing $\mathbf{w}_B$ accordingly, we write:

$$\mathbf{w}_B = \sum_{i=1}^{N} b_i \mathbf{x}_i + \mathbf{w}_{\perp},$$

where we note that the b_i coefficients could potentially be different.

We now proceed to prove that a non-zero component of $\mathbf{w}_{\perp}$ can only increase the value of the cost function. Proving this will lead us to our result, that an alternative vector $\mathbf{w}_B$ which does not lie on the span of the training set features cannot be a solution of our optimization problem.

Starting with the value of the empirical loss, we note that

$$\sum_{i=1}^{N} \max(0, 1 - y_i \mathbf{w}_B^T \mathbf{x}_i) = \sum_{i=1}^{N} \max(0, 1 - y_i (\sum_{i=1}^{N} b_i \mathbf{x}_i)^T \mathbf{x}_i), \tag{1}$$

since $\mathbf{x}_i \perp \mathbf{w}_{\perp} \forall i$: the component of the weight vector that is perpendicular to $\mathcal{X}$ cannot change the value of the empirical loss, since its inner product with any feature is zero, by definition. The empirical loss is thus only determined by the values of b_i.

Moving on to the value of the regularizer, by the Pythagorean theorem it follows that:

$$||\mathbf{w}_B|| = ||\sum_{i=1}^{N} b_i y_i \mathbf{x}_i|| + ||\mathbf{w}_{\perp}||$$

Combining these two results together, we have that our objective can be written as follows:

$$C(\mathbf{w}_B) = \sum_{i=1}^{N} \max(0, 1 - y_i (\sum_{i=1}^{N} b_i \mathbf{x}_i)^T \mathbf{x}_i) + \lambda \left(||\sum_{i=1}^{N} b_i \mathbf{x}_i|| + ||\mathbf{w}_{\perp}|| \right). \tag{2}$$

Putting things together, we have

$$\begin{aligned}
C(\mathbf{w}_{\mathcal{X}}) &= \sum_{i=1}^{N} \max(0, 1 - y_i (\sum_{i=1}^{N} a_i \mathbf{x}_i)^T \mathbf{x}_i) + \lambda ||\sum_{i=1}^{N} a_i \mathbf{x}_i|| &\quad (3)\\
&\leq \sum_{i=1}^{N} \max(0, 1 - y_i (\sum_{i=1}^{N} b_i \mathbf{x}_i)^T \mathbf{x}_i) + \lambda ||\sum_{i=1}^{N} b_i \mathbf{x}_i|| &\quad (4)\\
&\leq \sum_{i=1}^{N} \max(0, 1 - y_i (\sum_{i=1}^{N} b_i \mathbf{x}_i)^T \mathbf{x}_i) + \lambda \left(||\sum_{i=1}^{N} b_i \mathbf{x}_i|| + ||\mathbf{w}_{\perp}|| \right) &\quad (5)\\
&= C(\mathbf{w}_B) &\quad (6)
\end{aligned}$$

The first inequality follows from the definition of $\mathbf{w}_{\mathcal{X}}$, the second inequality from the fact that the norm of a vector is positive, while we proved the last equality above.

In conclusion, there cannot exist any value for $\mathbf{w}_B$ less than $\mathbf{w}_{\mathcal{X}}$.