Introduction to Supervised Learning, UCL, 2016-2017

Iasonas Kokkinos, i.kokkinoscs.ucl.ac.uk

Analytical Exercises (5 points/20, optional)

These are optional exercises that you can do in order to complement the points you will get by delivering your computing assignments.

As soon as your total (programming +analytical) credit goes over 5, the additional reward is split by two.

E.g. if you get 4/5 points from your programming assignments and 3/5 points from the analytical exercises you will not get 7 points; instead the 3 extra grades are split in 1+2, which gives you 4+1 (=5) + 2 /2 = 6 points.

Formula for the two grades being g_1, g_2 (note: corrected formula!):

$$g_1 + g_2 - 0.5 \max(g_1 + g2 - 5, 0)$$

Apart from helping you secure a good grade, solving these exercises will help you monitor your own understanding.

Note: certain of these exercises require more time, thought and effort than what we will expect in the exam.

Delivery: in person, in class (**not by email or moodle**), printed or handwritten document.

HARD deadline: 8AM, Wednesday 16 November.
We will spend the first hour of that lecture giving the solutions - anything delivered later than 8AM will be entirely ignored.

1 Section 1: Introduction, Probability (1 point)

1.1 Qualitative Understanding (0.33)

Suppose that you are working for the ministry of transportation. You want to design a machine learning algorithm that will predict the number of passengers that are anticipated to be using an underground line in the next hour. This could help e.g. plan and allocate personnel accordingly.

- What type of machine learning task does this correspond to? (clustering, dimensionality reduction, reinforcement learning, regression, classification, other)
- Propose three features (i.e. inputs to the function that you want to learn) that could be useful. Comment about whether these would be easy to measure or not.
- Propose two similar problems that could interest the department of transportation and would be phrased in similar terms. As another example that would be a valid answer, consider deciding whether there is an accident based on the density of cars on the street. Can you think of other problems?

Feel free to use any type of machine learning problem (classification, regression, etc).

Note: you may want to google internet of things and smart cities for ideas.

Reply in no more than 10 lines.

1.2 Probability (0.33)

Suppose we have three boxes r (red), b (blue), and g (green). Box r contains 30 apples, 4 oranges, and 3 limes, box b contains 1 apple, 1 orange and 0 limes, and box g contains 3 apples, 3 oranges, and 4 limes. Consider the following process: a box is chosen at random with probabilities $p(r) = 0.1, p(b) = 0.3, p(g) = 0.6$, and a piece of fruit is selected uniformly at random from the box.

- What is the probability that this process will select an apple?
- If we selected an orange, what is the probability that it came from the red box?

1.3 Probability (0.33)

Suppose we have two six-sided dice, one is fair and the other always comes up 6. We will toss a biased coin to decide which die to roll. The coin comes up heads with probability p. If the coin turns up heads we will toss the fair die and if the coin turns up tails we will toss the other. Let x be the value of the die roll obtained under this process.

- What is the expectation of x?
- What is the variance of x?

2 Section 2: Linear Regression (1.5 points)

2.1 Qualitative understanding (0.33)

Suppose that your task is to estimate for how much a second-hand Mini Cooper be sold. Your inputs are

- the year in which the car was manufactured
- the number of miles it has run
- its condition, as estimated by the car owner on a scale from 1 to 4, with 4 being excellent
- whether it has been involved in a car accident, as indicated by the cars insurance record, with 0 being no accident and 1 being with accident.

You represent these inputs as a four-dimensional vector, $\mathbf{x} = (y, m, c, a)$, for year, miles, condition, accident.

Given previous recordings of sales for Mini Cooper models, you fit a linear regression model. Your price gets estimated by a linear function of the form $y = \mathbf{w}^T\mathbf{x} + c$, where y is measured in thousands of pounds, and c represents a constant term that is estimated jointly with $\mathbf{w}$.

- Among the two options, $\mathbf{w} = (-1, -2, 1, -10)$ and $\mathbf{w} = (-1, -2, 10, -1)$ which one seems more reasonable to you? Why?
- If you have 3 previous sales records of Mini Cooper cars, would you trust your regression results, and why? Can you mathematically justify your answer?

Now consider that you have a big dataset with one thousand sales records of Mini Cooper cars. Having fitted a linear model you are dissatisfied with your results, and fear that the linearity assumption is not valid - and therefore costing you money. You turn to a nonlinear regression approach, by using a nonlinear embedding function $\phi(\mathbf{x})$that maps $\mathbf{x}$ from $\mathbb{R}^4$ into $\mathbb{R}^{10000}$. This ends up having zero training error.

How will you convince yourself that this is a good result? If it turns out to be a bad idea, how will you improve your approach?

Reply in no more than 15 lines.

2.2 Least Squares Estimation (0.33)

In the beginning of the slides for 'Week2(b)' we give a simple proof for the least squares fitting of a line to data, by using the expression for the derivative of the sum-of-squared-error loss with respect to the individual parameters.

Then for the D-dimensional case we somehow make a jump, and obtain the expression for the solution using the vector derivative of the algebraic formulation of the loss.

Derive the solution to the general, D-dimensional case while using similar steps to the ones used for the problem of fitting a line to the data.

2.3 Regularization and priors (0.5)

Let $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ be a function defined by a vector of parameters $\mathbf{w}$. In class we showed that least squares regression with a training set $S = \{(x^1, y^1), ..., (x^N, y^N)\}$ corresponds to the maximum likelihood estimate of $\mathbf{w}$ assuming

- $y^i = f_{\mathbf{w}}(\mathbf{x}^i) + e^i$
- the errors e^i are independent and distributed according to a Normal distribution with mean $\mathbf{0}$ and standard deviation σ^2.

Now suppose we have a prior distribtion $p(\mathbf{w})$ over the parameters $\mathbf{w}$, defined by a multivariate Normal with mean $\boldsymbol{\mu} = \mathbf{0}$ and covariance matrix aI for some $a \in R$ and I being the identity matrix.

The Maximum a posteriori (MAP) estimate of $\mathbf{w}$ is the vector maximizing the posterior probability of $\mathbf{w}$ given S, defined as: $\mathbf{w}_{MAP} \doteq \arg\max_{\mathbf{w}} p(\mathbf{w}|S)$.

- Show that $\mathbf{w}_{MAP} = \arg\max_{\mathbf{w}} p(S|\mathbf{w})p(\mathbf{w})$.
- Show that there exists a λ such that:
  $\mathbf{w}_{MAP} = \arg\min_{\mathbf{w}} \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{N}(f_{\mathbf{w}}(\mathbf{x}^i) - y^i)2$
- What is the relationship between λ, a and σ^2?

2.4 Linear Discriminants for multiple classes (0.33)

A K-class discriminant is obtained by training K linear classifiers of the form $f_k(\mathbf{x}) = \mathbf{w}_k^T\mathbf{x}+b_k$ and assigning a point to class C_k if $f_k(\mathbf{x}) > f_j(\mathbf{x})$ for all $j \neq k$.

- Write the equation of the hyperplane separating class j and k.
- If $\mathbf{x}_A$ and $\mathbf{x}_B$ are both classified as class j, then show that any point on the line $\mathbf{x} = \lambda\mathbf{x}_A + (1-\lambda)\mathbf{x}_B$ where $0 \leq \lambda \leq 1$, is also classified as class j.

3 Section 3: Logistic Regression (1 point)

3.1 Optimization for Logistic Regression (0.66)

The cost function for logistic regression covered in class was given by the following expression:

$$L(\mathbf{w}) = -\sum_{i=1}^{N} y^i \log(g(\langle \mathbf{x}^i, \mathbf{w}\rangle) + (1-y^i)\log(1-g(\langle \mathbf{x}^i, \mathbf{w}\rangle)). \tag{1}$$

During class we came up with the expressions for the gradient vector $\nabla L(\mathbf{w})$ and the Hessian matrix $H(\mathbf{w})$.

- In the Week 4 Lecture you can find two expressions for the Hessian matrix of the cost function. The first is expressing the (k, j) element of this matrix in terms of a summation that runs over the training set samples, $i = 1, \ldots, N$:

$$H_{k,j}(\mathbf{w}) = \frac{\partial^2 L(\mathbf{w})}{\partial w_k \partial w_j} = \sum_{i=1}^{N} \mathbf{x}_k^i g(\mathbf{w}^T\mathbf{x}^i)(1-g(\mathbf{w}^T\mathbf{x}^i)\mathbf{x}_j^I$$

  The second is expressing the whole Hessian matrix in terms of matrix operations:

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

  where $\mathbf{X}$ is the matrix introduced in Lecture 2:

$$\mathbf{X} = \left[\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^N\right]^T$$

  and $\mathbf{R}$ is a diagonal matrix,

$$\mathbf{R} = \begin{bmatrix} R_{1,1} & 0 & \ldots & 0 \\ 0 & R_{2,2} & \ldots & 0 \\ \vdots & & & \\ 0 & 0 & \ldots & R_{N,N} \end{bmatrix}$$

  with diagonal elements given by $R_{i,i} = g(\mathbf{w}^T\mathbf{x}^i)(1-g(\mathbf{w}^T\mathbf{x}^i)$. Show that the two expressions are the same.

- Find how the $\nabla L(\mathbf{w})$ and $H(\mathbf{w})$ would change if one adds an l_2 regularization term to the objective:

$$L(\mathbf{w}) \quad = \quad \lambda\|\mathbf{w}\|_2^2 - \sum_{i=1}^{N} y^i \log(g(\langle \mathbf{x}^i, \mathbf{w}\rangle) + (1-y^i)\log(1-g(\langle \mathbf{x}^i, \mathbf{w}\rangle)), \text{ where } \|\mathbf{w}\|_2^2 = \sum_j \mathbf{w}_j^2 \tag{2}$$

3.2 Multi-class classification and logistic regression (0.33 points)

For the two-class classification problem we had worked with the following expression for the class posterior:

$$P(y = 1|\mathbf{x}; \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x})}$$

Then, when talking about multi-class classification, with C classes, we introduced the softmax operation and said that the class posteriors is given by the following expression:

$$P(y = k|\mathbf{x}; \mathbf{W}) = \frac{\exp(\mathbf{w}_k^T\mathbf{x})}{\sum_{c=1}^{C} \exp(\mathbf{w}_c^T\mathbf{x})}, \quad \mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_C]$$

Show that when $C = 2$ the two expressions are equivalent, and recover the relationship between $\mathbf{w}$ in Eq. 3 and $\mathbf{w}_1, \mathbf{w}_2$ in Eq. 3.

4 Section 4: SVMs (1.5 points)

4.1 Geometry (.5)

Consider a linear discriminant $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$ with a decision boundary defined by $\mathbf{w}^T\mathbf{x} + b = 0$. Prove that the direction of the weight vector $\mathbf{w}$ is perpendicular to the decision boundary.

4.2 Representer Theorem (1)

In the course on SVMs we said that the form of the weight vector that results from the solution of the SVM optimization problem is of the form $\mathbf{w}^* = \sum_{i=1}^{N} a^i y^i \mathbf{x}^i$, namely lies on the span of the basis defined by the training features.

Prove that this is true. Hint: use reduction to the absurd - express an alternative candidate solution as $\mathbf{w}^a = \sum_{i=1}^{N} b^i y^i \mathbf{x}^i + \mathbf{w}_\perp$ where $\mathbf{w}_\perp^T \mathbf{x}^i = 0, \forall i$. Show that this cannot be any better than $\mathbf{w}^*$. Use the Pythagorean theorem to express the norm of $\mathbf{w}^a$ in terms of the sum of the norms of the vectors that lie on the span of the training features and $\mathbf{w}_\perp$.