Machine Learning for Finance – Problem Set 3

Neal Parikh

February 27, 2018

Instructions. Do not refer to any outside sources to complete this assignment, in accordance with the honor code. If you discussed any problems with other students, indicate that in your solutions.

1. *Weighted least squares.* In least squares, the objective to be minimized is

$$\|Xw - y\|_2^2 = \sum_{i=1}^{N} (w^T x_i - y_i)^2,$$

where x_i^T are the rows of X and the model parameters $w \in \mathbf{R}^n$ is the optimization variable. In *weighted least squares*, we instead minimize the objective

$$\sum_{i=1}^{m} \lambda_i (w^T x_i - y_i)^2,$$

where λ_i are fixed positive weights. The weights allow assigning a different amount of emphasis on different components of the residual vector.

(a) Show that the weighted least squares objective can be expressed as $\|D(Xw - y)\|_2^2$ for an appropriate diagonal matrix D. This allows solving the weighted least squares problem as a standard least squares problem by minimizing $\|Uw - v\|_2^2$, where $U = DX$ and $v = Dy$.

(b) Show that when X has linearly independent columns, so does U.

(c) The least squares approximate solution is given by $\hat{w} = (X^T X)^{-1} X^T y$. Give a similar formula for the solution of the weighted least squares problem.

Hint. You may want to use the matrix $\Lambda = \mathbf{diag}(\lambda)$ in your formula.

(d) Consider a training set of N independent examples (x_i, y_i) in which the y_i's were observed with differing variances. Specifically, suppose that

$$p(y_i \mid x_i; w) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma_i^2} \right),$$

i.e., $y_i \mid x_i \sim \mathrm{N}(w^T x_i, \sigma_i^2)$, where the σ_i are fixed and known constants. Show that finding the maximum likelihood estimate of w reduces to solving a weighted linear regression problem. State what the λ_i are in terms of the σ_i.

(e) *Locally weighted linear regression.* Suppose we want to predict the output y^{new} for a new query point x^{new}. Consider using the weights

$$\lambda_i = \exp\left(\frac{-(x^{\text{new}} - x_i)^2}{2\tau^2}\right),$$

where $\tau > 0$ is a fixed *bandwidth parameter*. Explain in English what this model is doing. Comment on how its behavior varies with different values of τ, and especially on what happens to the fit when τ is very large or very small.

2. *Data matrix in autoregressive time series model.* Suppose that $z_1, z_2, \ldots$ is a time series. An *autoregressive model* for the time series has the form

$$\hat{z}_{t+1} = w_1 z_t + \cdots + w_M z_{t-M+1}, \quad t = M, M+1, \ldots$$

where M is the *memory* or *lag* of the model. An autoregressive model is also referred to as an *AR model*, or an *AR(M) model* for a particular memory M. Here $\hat{z}_{t+1}$ is the prediction of z_{t+1} made at time t (when $z_t, \ldots, z_{t-M+1}$ are known). This prediction is a linear function of the previous M values of the time series. With a good choice of model parameters, the AR model can be used to predict the next value in a time series, given the current and previous M values. This has many practical uses.

We can use least squares or linear regression to fit the parameters of an AR(M) model based on the observed data $z_1, \ldots, z_T$ by minimizing the sum of squares of the *prediction errors* $z_{t+1} - \hat{z}_{t+1}$ over $t = M, \ldots, T-1$, *i.e.*,

$$(z_{M+1} - \hat{z}_{M+1})^2 + \cdots + (z_T - \hat{z}_T)^2.$$

(We must start the predictions at $t = M$, since each prediction involves the previous M time series values, and we do not know $z_0, z_{-1}, \ldots$.)

Find the matrix X and vector y for which $\|Xw - y\|_2^2$ gives the sum of the squares of the prediction errors. Show that X is a Toeplitz matrix, *i.e.*, that entries X_{ij} with the same value of $i - j$ are the same. Indicate how many features and examples are in the regression.

3. *Augmenting features with the average.* You are fitting a regression model $\hat{y} = x^T\beta + v$ to data, computing the model coefficients β and v using least squares. A friend suggests adding a new feature, which is the average of the original features. (That is, he suggests using the new feature vector $\tilde{x} = (x, \mathbf{avg}(x))$.) He explains that by adding this new feature, you might end up with a better model. Is this a good idea?

4. *Sigmoid function.* Recall that the sigmoid function is

$$s(x) = 1/(1 + e^{-x}).$$

(a) Show that its derivative satisfies the property

$$s'(x) = s(x)(1 - s(x)).$$

(b) Show that if

$$\log \frac{p(y = 1 \mid x)}{p(y = 0 \mid x)} = w^T x,$$

where $x \in \mathbf{R}^n$ and $y \in \{0, 1\}$, then

$$p(y = 1 \mid x) = s(w^T x).$$

5. *Convexity of logistic regression.* Consider the log-likelihood function for logistic regression:

$$\ell(w) = \sum_{i=1}^{N} y_i \log s(w^T x_i) + (1 - y_i) \log(1 - s(w^T x_i)),$$

where s is the sigmoid function. Show that ℓ is concave in w.

6. *Generalized linear models for count and rate data.* Consider a process in which events occur independently and continuously at a constant rate. (Though not relevant to the problem, such a process is known as a *Poisson process.*) It is used to model a wide variety of situations, such as the arrival of customers at a store or incoming messages at an exchange; it can also be used to model spatial data like locations of trees in a forest or meteor strikes of Earth.

Rather than working with the process itself, we are often interested in two particular aspects of such processes. The *number of events* occurring in a fixed interval follows a Poisson distribution, a discrete distribution given by the mass function

$$p(z) = \frac{e^{-\lambda} \lambda^z}{z!},$$

with parameter $\lambda > 0$, where z is a nonnegative integer. The parameter is often known as a *rate parameter* and is also the mean of the distribution. It is commonly used to model count or rate data, such as the number of patients arriving to a hospital during business hours.

The *time between events* occurring is described by the *exponential distribution* (not to be confused with the exponential family), a continuous distribution given by the density function

$$p(z) = \lambda \exp(-\lambda z),$$

with parameter $\lambda > 0$, where $z \in \mathbf{R}_+$. Here, λ is also called a rate parameter, but the mean of the distribution is $1/\lambda$. For example, if messages arrive independently at random with one every s seconds on average, then the distribution of how long one must wait for a message is exponential with mean s. The distribution often arises in waiting problems and can be used to model, *e.g.*, service times of agents in a system or the time until default or payment to debtholders when modeling credit risk.

Generalized linear models with a Poisson or exponential response are also closely connected to survival analysis and reliability engineering.

(a) Show that the Poisson distribution is in the exponential family. What is the canonical response function for a GLM with a Poisson response (known as *Poisson regression*)?

(b) Show that the exponential distribution is in the exponential family. What is the canonical response function for a GLM with an exponential response?

7. *Maximum likelihood estimation and moment matching.* Suppose we have a random variable following the exponential family

$$p(x; \theta) = \exp\left(\theta^T \varphi(x) - A(\theta)\right),$$

where $\varphi(x) = (\varphi_1(x), \ldots, \varphi_K(x))$ is given and the parameters $\theta \in \mathbf{R}^K$ are unknown.

   (a) Given a dataset $x_1, \ldots, x_N$, show that the maximum likelihood estimate $\hat{\theta}$ of the parameters satisfy the condition

   $$\sum_x p(x; \hat{\theta}) \varphi_k(x) = \frac{1}{N} \sum_{i=1}^{N} \varphi_k(x_i)$$

   for all k, where the lefthand sum is over all x in the support of the distribution and the righthand sum is over all data points. A shorthand for this result is that the 'model average', the expected value under the fitted model, of each sufficient statistic φ_k must equal the empirical average of the sufficient statistic found in the training data. These are sometimes called the *moment matching conditions*. (We could also include $h(x)$ in the density above; it only slightly clutters the derivation.)

   Hint. Differentiate the log-likelihood and rearrange.

   (b) Suppose the distribution in question is the multinomial distribution and the sufficient statistics are indicator functions of each outcome. Give an interpretation for the moment matching conditions in this setting and briefly discuss its implications.

8. *Kullback-Leibler divergence and maximum likelihood.* The *Kullback-Leibler divergence*, also called KL divergence and *relative entropy*, between two discrete-valued distributions p and q is given by

$$\mathrm{KL}(p \,\|\, q) = \sum_x p(x) \log \frac{p(x)}{q(x)}.$$

(Here, it is assumed that $q(z) = 0$ implies $p(z) = 0$, and that $0 \log 0 = 0$.) The KL divergence is also often denoted $D(p \,\|\, q)$.

The KL divergence is a measure of the 'distance' between two probability distributions and has many interpretations. (We use 'distance' in quotes because it is not a metric. In particular, it is not symmetric, so care must be taken to indicate which of $\mathrm{KL}(p \,\|\, q)$ or $\mathrm{KL}(q \,\|\, p)$ is meant in a given situation.) Roughly, $\mathrm{KL}(p \,\|\, q)$ is a measure of the inefficiency of assuming that the distribution is q (which is often a model or approximation) when the true distribution is p.

   (a) *Nonnegativity.* Prove that $\mathrm{KL}(p \,\|\, q) \geq 0$, and that $\mathrm{KL}(p \,\|\, q) = 0$ if and only if $p = q$. This is known as *Gibbs' inequality* or the *information inequality.*

   Hint. Use Jensen's inequality: For any random variable z, $f(\mathrm{E}[z]) \leq \mathrm{E}[f(z)]$ when f is convex, with equality either when f is not strictly convex or when z is a constant, *i.e.*, $z = \mathrm{E}[z]$ with probability 1.

   (b) *Chain rule.* The KL divergence between two conditional distributions is given by

   $$\mathrm{KL}(p(x \,|\, y) \,\|\, q(x \,|\, y)) = \sum_y p(y) \left(\sum_x p(x \,|\, y) \log \frac{p(x \,|\, y)}{q(x \,|\, y)} \right).$$

Prove that

$$\mathrm{KL}(p(x,y) \,\|\, q(x,y)) = \mathrm{KL}(p(x) \,\|\, q(x)) + \mathrm{KL}(p(y \,|\, x) \,\|\, q(y \,|\, x)).$$

(c) *KL divergence and maximum likelihood.* Suppose we are given a training set $\{x_1, \ldots, x_N\}$ and let the empirical distribution be

$$\tilde{p}(x) = \frac{1}{N} \sum_{i=1}^{N} [x_i = x].$$

Let $\mathcal{P} = \{p_\theta \mid \theta \in \Theta\}$ be a family of probability distributions indexed by a parameter θ. Show that finding the maximum likelihood estimate of θ is equivalent to finding the $p_\theta \in \mathcal{P}$ with minimal KL divergence from the empirical distribution $\tilde{p}$, *i.e.*, that

$$\operatorname*{argmin}_{\theta \in \Theta} \mathrm{KL}(\tilde{p} \,\|\, p_\theta) = \operatorname*{argmax}_{\theta \in \Theta} \left(\sum_{i=1}^{N} \log p_\theta(x_i) \right).$$

This is sometimes referred to as the *M-projection* of $\tilde{p}$ onto $\mathcal{P}$.