Machine Learning for Finance – Problem Set 3 Solutions

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

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

   i.e., $y_i \,|\, x_i \sim \mathrm{N}(w^T x_i, \sigma_i^2)$, 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.

Solution.

(a) Since the weights are positive, we can write the objective as

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

where D is the diagonal matrix

$$D = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_m} \end{bmatrix}.$$

(b) We show that $Uw = 0$ implies $w = 0$.

Suppose $Uw = DXw = 0$. Then $Xw = 0$ because D is a diagonal matrix with positive diagonal entries, and hence invertible. By assumption, X has linearly independent columns, so $Xw = 0$ implies $w = 0$.

(c) The solution of the weighted least squares problem is

$$\begin{aligned} (U^T U)^{-1} U^T v &= ((DX)^T (DX))^{-1} (DX)^T (Dy) \\ &= (X^T D^2 X)^{-1} (X^T D^2 y) \\ &= (X^T \Lambda X)^{-1} (X^T \Lambda y), \end{aligned}$$

where $\Lambda = D^2 = \mathbf{diag}(\lambda)$.

(d) Maximum likelihood estimation involves maximizing

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

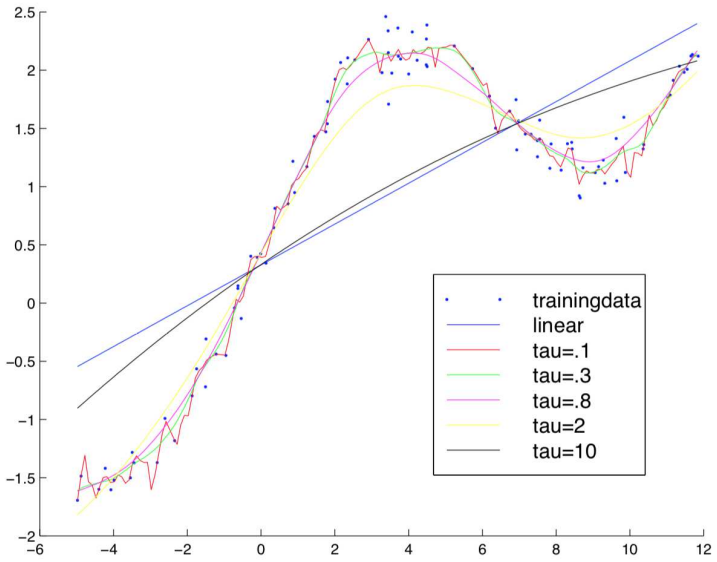
which can be seen to be equivalent to minimizing

$$\frac{1}{2} \sum_{i=1}^{N} \frac{1}{\sigma_i^2} (y_i - w^T x_i)^2,$$

so $\lambda_i = 1/\sigma_i^2$.

(e) At a given point x^{new}, locally weighted regression fits a linear regression model, but places much more emphasis on points close to x^{new}; points far away have much less weight and do not affect the fit much. The size of the 'neighborhood' considered in the local fit is controlled by τ, with larger values of τ placing more weight on points farther away from the query point. As $\tau \to \infty$, the weights $\lambda_i \to 1$, so in the limit this approaches the standard linear regression model.

The fit obtained by this method is *not* linear, and the curve through the points will be more or less 'wiggly' depending on how small or large the bandwidth parameter τ is:



(This diagram is due to Andrew Ng.) The smaller the bandwidth, the fewer training samples are effectively taken into account in the regression, so the regression results become susceptible to noise in those few training samples. For larger τ, there are enough training samples to more reliably fit straight lines; unfortunately, a straight line is not the right model for this data, so we also get a bad fit for large values of τ.

Locally weighted linear regression is an example of a *nonparametric* regression method; it can be viewed as a blend of ideas from linear regression and k-nearest neighbors. The (unweighted) linear regression model is known as a *parametric* learning algorithm, because it has a fixed, finite number of parameters which are fit to the data. Once the parameters are fit and stored, the training data is no longer needed to make future predictions. In contrast, to make predictions using locally weighted linear regression, we need to keep the entire training set around. The term nonparametric (roughly) refers to the fact that the amount of information we need to keep in order to represent the model grows linearly with the size of the training set.

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.

Solution.

$$X = \begin{bmatrix} z_M & z_{M-1} & z_{M-2} & \cdots & z_1 \\ z_{M+1} & z_M & z_{M-1} & \cdots & z_2 \\ z_{M+2} & z_{M+1} & z_M & \cdots & z_2 \\ \vdots & \vdots & \vdots & & \vdots \\ z_{T-1} & z_{T-2} & z_{T-3} & \cdots & z_{T-M} \end{bmatrix}, \qquad y = \begin{bmatrix} z_{M+1} \\ z_{M+2} \\ z_{M+3} \\ \vdots \\ z_T \end{bmatrix},$$

so there are $T - M$ examples and M features.

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?

   Solution. This is a bad idea. With the new feature, the data matrix has a new column that is $1/n$ times the sum of the other columns. In particular, the new column is a linear combination of the other columns. It follows that the data matrix does not have linearly independent columns, so we cannot solve the associated least squares problem.

   Even if we could solve the associated least squares problem, our results would not be any better. Any linear combination of the columns of the new data matrix is also a linear combination of the columns of the original data matrix, and it follows that we cannot find a linear combination of the columns of the new matrix that is closer to b than the optimal linear combination of the columns of the original matrix.

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

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

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

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

Solution.

(a) This is a straightforward calculation:

$$\begin{aligned} s'(z) &= \frac{d}{dz} \frac{1}{1 + e^{-z}} \\ &= \frac{1}{(1 + e^{-z})^2} \cdot e^{-z} \\ &= \frac{1}{1 + e^{-z}} \cdot \left(1 - \frac{1}{1 + e^{-z}}\right) \\ &= s(z)(1 - s(z)). \end{aligned}$$

This property can be useful in derivations.

(b) To lighten notation, let $q = p(y = 1 \,|\, x)$, so $1 - q = p(y = 0 \,|\, x)$. It is easy to rearrange

$$\log \frac{q}{1 - q} = w^T x$$

as

$$q = \frac{\exp(w^T x)}{1 + \exp(w^T x)}.$$

The result follows from observing that

$$s(z) = \frac{1}{1 + e^{-z}} = \frac{e^z}{1 + e^z}.$$

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.

Solution. Let $f(x) = s(w^T x)$. From $s'(z) = s(z)(1 - s(z))$, we have that

$$\nabla_w f(x) = f(x)(1 - f(x))x.$$

The gradient of the log-likelihood is given by

$$\nabla \ell(w) = \sum_{i=1}^{N} (y_i - f(x_i)) x_i.$$

To see this, suppose $N = 1$ (so there is a single example (x, y)) and consider taking the partial derivative of ℓ with respect to w_k:

$$\begin{aligned}
\frac{\partial}{\partial w_k} \ell(w) &= \left(\frac{y}{f(x)} - \frac{1-y}{1-f(x)} \right) \frac{\partial}{\partial w_k} f(x) \\
&= \left(\frac{y}{f(x)} - \frac{1-y}{1-f(x)} \right) f(x)(1-f(x)) x^k \\
&= \left(y(1-f(x)) - (1-y) f(x) \right) x^k \\
&= (y - f(x)) x^k,
\end{aligned}$$

where x^k is the kth feature of x.

Taking partial derivatives again, entries of the Hessian are given by

$$(\nabla^2 \ell(w))_{kl} = -\sum_{i=1}^{N} f(x_i)(1-f(x_i)) x_i^k x_i^l,$$

so

$$\nabla^2 \ell(w) = -\sum_{i=1}^{N} f(x_i)(1-f(x_i)) x_i x_i^T.$$

That $\nabla^2 \ell(w) \preceq 0$ follows from $x_i x_i^T \succeq 0$ (see problem set 1) and $s(z) \in (0, 1)$.

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?

Solution.

(a) The distribution can be rewritten as

$$\begin{aligned} p(z) &= \frac{e^{-\lambda} \exp \log \lambda^z}{z!} \\ &= \frac{1}{z!} \exp(z \log \lambda - \lambda), \end{aligned}$$

so we have

$$h(z) = 1/z!, \quad \theta = \log \lambda, \quad A(\theta) = e^\theta.$$

The canonical link function is the logarithm, so the response function is $\mathrm{E}[z] = \lambda = e^\theta$.

(b) Here, we only need the simplest version of an exponential family density. The density can be written as

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

so

$$\theta = -\lambda, \quad A(\eta) = -\log(-\theta).$$

The canonical response function is

$$\mathrm{E}[z] = \frac{1}{\lambda} = -\frac{1}{\theta}.$$

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.

Solution.

(a) The log-likelihood is

$$\ell(\theta) = -NA(\theta) + \sum_{i=1}^{N} \theta^T \varphi(x_i),$$

so

$$\frac{\partial}{\partial \theta_k} \ell(w) = -N \frac{\partial}{\partial \theta_k} A(\theta) + \sum_{i=1}^{N} \varphi_k(x_i).$$

Recall that

$$A(\theta) = \log Z(\theta) = \log \sum_x \exp(\theta^T \varphi(x)).$$

Then

$$\begin{aligned} \frac{\partial}{\partial \theta_k} A(\theta) &= \frac{1}{Z(\theta)} \sum_x \frac{\partial}{\partial \theta_k} \exp(\theta^T \varphi(x)) \\ &= \frac{1}{Z(\theta)} \sum_x \exp(\theta^T \varphi(x)) \varphi_k(x) \\ &= \sum_x p(x; \theta) \varphi_k(x). \end{aligned}$$

It is worth pausing at this result, since it is of general significance: This gives that $\nabla A(\theta) = \mathrm{E}[\varphi(x)]$, *i.e.*, that the gradient of the log partition function is the expected value of the sufficient statistics.

This gives that

$$\frac{\partial}{\partial \theta_k} \ell(\theta) = -N \sum_x p(x; \theta) \varphi_k(x) + \sum_{i=1}^{N} \varphi_k(x_i).$$

At the maximum likelihood estimate $\hat{\theta}$ (obtained by setting the previous expression to zero and solving), we have that

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

for each k. This could be written more compactly as $\mathrm{E}_{\hat{\theta}}[\varphi_k(x)] = \overline{\varphi_k(x)}$.

(b) The expected value of the indicator function of an event is the probability of the event. Moreover, the probabilities of each possible outcome occurring are precisely the 'usual' parameters of the distribution, so this result provides a very simple way to estimate these parameters, often with computation no more involved than counting or averaging. If we are modeling a die roll and we see 23 occurrences of a 3 appearing in 100 rolls, we simply set the parameter $\theta_3 = 0.23$; this similarly applies to spam filtering or other examples in natural language processing.

Of course, this also applies to many other distributions. To estimate the mean parameter of a Gaussian, we simply use the average of the observed sample points.

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 unusual notation is common in information theory.

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 $\mathcal{P}$; that is, we can view this geometrically as carrying out a nonlinear projection (with distance measured by the KL divergence) of the point $\tilde{p}$ onto the set $\mathcal{P}$. This perspective is studied in great depth in the field of *information geometry*.

Solution.

(a) The result follows from applying Jensen's inequality with the negative logarithm (a strictly convex function) as f and q/p as the random variable:

$$\begin{aligned}
-\mathrm{KL}(p \,\|\, q) &= -\sum_x p(x) \log \frac{p(x)}{q(x)} \\
&= \sum_x p(x) \log \frac{q(x)}{p(x)} \\
&\leq \log \sum_x p(x) \frac{q(x)}{p(x)} \\
&= \log \sum_x q(x) \\
&= \log 1 \\
&= 0.
\end{aligned}$$

Since the negative logarithm is strictly convex, it follows that $\mathrm{E}[q/p] = q/p$. But $\mathrm{E}[q/p] = \sum_x q(x) = 1$, so $p(x) = q(x)$ for all x, giving $\mathrm{KL}(p \,\|\, q) = 0$ if and only if $p = q$.

(We show that $-\mathrm{KL}(p \,\|\, q) \leq 0$, which may seem odd, rather than $\mathrm{KL}(p \,\|\, q) \geq 0$ directly, only to apply Jensen's inequality as stated above more directly. There are a number of other ways to go through the same argument that you may find more intuitive.)

(b)

$$\begin{aligned}
\mathrm{KL}(p(x,y) \,\|\, q(x,y)) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(x)p(y \,|\, x)}{q(x)q(y \,|\, x)} \\
&= \sum_{x,y} \left(p(x,y) \log \frac{p(x)}{q(x)} + p(x,y) \log \frac{p(y \,|\, x)}{q(y \,|\, x)} \right) \\
&= \sum_{x,y} p(x,y) \log \frac{p(x)}{q(x)} + \sum_{x,y} p(x)p(y \,|\, x) \log \frac{p(y \,|\, x)}{q(y \,|\, x)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(x)}{q(x)} + \sum_{x} p(x) \sum_{y} p(y \,|\, x) \log \frac{p(y \,|\, x)}{q(y \,|\, x)} \\
&= \mathrm{KL}(p(x) \,\|\, q(x)) + \mathrm{KL}(p(y \,|\, x) \,\|\, q(y \,|\, x)).
\end{aligned}$$

(c)

$$\begin{aligned}
\operatorname{argmin} \mathrm{KL}(\tilde{p} \,\|\, p_\theta) &= \operatorname{argmin} \left(\sum_x \tilde{p}(x) \log \tilde{p}(x) - \tilde{p}(x) \log p_\theta(x) \right) \\
&= \operatorname{argmax} \left(\sum_x \tilde{p}(x) \log p_\theta(x) \right) \\
&= \operatorname{argmax} \left(\sum_x \frac{1}{N} \sum_{i=1}^{N} [x_i = x] \log p_\theta(x) \right) \\
&= \operatorname{argmax} \frac{1}{N} \left(\sum_{i=1}^{N} \sum_x [x_i = x] \log p_\theta(x) \right) \\
&= \operatorname{argmax} \frac{1}{N} \left(\sum_{i=1}^{N} \log p_\theta(x_i) \right) \\
&= \operatorname{argmax} \left(\sum_{i=1}^{N} \log p_\theta(x_i) \right).
\end{aligned}$$