CIS 520: Machine Learning **Spring 2018: Lecture 4**

Least Squares Regression

Lecturer: Shivani Agarwal

Outline

- Regression and conditional expectation
- Linear least squares regression
- Ridge regression and Lasso
- Probabilistic view

1 Regression and Conditional Expectation

In this lecture we consider regression problems, where there is an instance space $\mathcal{X}$ as before, but labels and predictions are real-valued: $\mathcal{Y} = \widehat{\mathcal{Y}} = \mathbb{R}$ (such as in a weather forecasting problem, where instances might be satellite images showing water vapor in some region and labels/predictions might be the amount of rainfall in the coming week, or in a stock price prediction problem, where instances might be feature vectors describing properties of stocks and labels/predictions might be the stock price after some time period). Here one is given a training sample $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathbb{R})^m$, and the goal is to learn from S a **regression model** $f_S : \mathcal{X} \to \mathbb{R}$ that predicts accurately labels of new instances in $\mathcal{X}$.

What should count as a good regression model? In other words, how should we measure the performance of a regression model? A widely used performance measure involves the **squared loss function**, $\ell_{\text{sq}} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, defined as

$$\ell_{\text{sq}}(y, \widehat{y}) = (\widehat{y} - y)^2 \,.$$

The loss of a model $f : \mathcal{X} \to \mathbb{R}$ on an example (x, y) is measured by $\ell_{\text{sq}}(y, f(x)) = (f(x) - y)^2$. Assuming examples are drawn from some joint probability distribution D on $\mathcal{X} \times \mathbb{R}$, the **squared-loss generalization error of** $f : \mathcal{X} \to \mathbb{R}$ **w.r.t.** D is then given by

$$\text{er}_D^{\text{sq}}[f] = \mathbf{E}_{(X,Y)\sim D}\left[(f(X) - Y)^2\right] .$$

What would be the optimal regression model for D under the above loss? We have,

$$\text{er}_D^{\text{sq}}[f] = \mathbf{E}_X\left[\mathbf{E}_{Y|X}\left[(f(X) - Y)^2\right]\right] .$$

Now, for each x, we know (and it is easy to see) that the value c minimizing $\mathbf{E}_{Y|X=x}[(c - Y)^2]$ is given by $c^* = \mathbf{E}[Y|X = x]$. Therefore the optimal regression model is simply the **conditional expectation function**, also called the **regression function of** Y **on** x:

$$f^*(x) = \mathbf{E}[Y|X = x] \,.$$

The conditional expectation function plays the same role for regression w.r.t. squared loss as does a Bayes optimal classifier for binary classification w.r.t. 0-1 loss. The minimum achievable squared error w.r.t. D is simply

$$\mathrm{er}_D^{\mathrm{sq},*} = \inf_{f:\mathcal{X}\to\mathbb{R}} \mathrm{er}_D^{\mathrm{sq}}[f] = \mathrm{er}_D^{\mathrm{sq}}[f^*] = \mathbf{E}_X\left[\mathbf{E}_{Y|X}\left[(Y - \mathbf{E}[Y|X])^2\right]\right] = \mathbf{E}_X\left[\mathbf{Var}[Y|X]\right],$$

which is simply the expectation over X of the conditional variance of Y given X; this plays the same role as the Bayes error for 0-1 binary classification.

2 Linear Least Squares Regression

For the remainder of the lecture, let $\mathcal{X} = \mathbb{R}^d$, and let $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)) \in (\mathcal{X} \times \mathbb{R})^m$. We start with a simple approach which does not make any assumptions about the underlying probability distribution, but simply fits a linear regression model of the form $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ to the data by minimizing the empirical squared error on S, $\widehat{\mathrm{er}}_S^{\mathrm{sq}}[f_{\mathbf{w}}] = \frac{1}{m}\sum_{i=1}^m (f_{\mathbf{w}}(\mathbf{x}_i) - y_i)^2$:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{m}\sum_{i=1}^m \left(\mathbf{w}^\top \mathbf{x}_i - y_i\right)^2. \tag{1}$$

Setting the gradient of the above objective to zero yields

$$\frac{2}{m}\sum_{i=1}^m \left(\mathbf{w}^\top \mathbf{x}_i - y_i\right)\mathbf{x}_i = \mathbf{0}.$$

We can rewrite this using matrix notation as follows: let

$$\mathbf{X} = \begin{bmatrix} -\ \mathbf{x}_1^\top\ - \\ -\ \mathbf{x}_2^\top\ - \\ \vdots \\ -\ \mathbf{x}_m^\top\ - \end{bmatrix} \in \mathbb{R}^{m\times d} \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} \in \mathbb{R}^m;$$

then we have

$$\mathbf{X}^\top\mathbf{X}\mathbf{w} - \mathbf{X}^\top\mathbf{y} = \mathbf{0}.$$

These are known as the **normal equations** for least squares regression and yield the following solution for $\mathbf{w}$ (assuming $\mathbf{X}^\top\mathbf{X}$ is non-singular):

$$\widehat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

The **linear least squares regression** model is then given by

$$f_S(\mathbf{x}) = \widehat{\mathbf{w}}^\top\mathbf{x}.$$

The solution $\widehat{\mathbf{w}}$ can be viewed as performing an orthogonal projection of the label vector $\mathbf{y}$ in $\mathbb{R}^m$ onto the d-dimensional subspace (assuming $m > d$) spanned by the d vectors $\tilde{\mathbf{x}}_k = (x_{1k}, \ldots, x_{mk})^\top \in \mathbb{R}^m$, $k = 1, \ldots, d$ (in particular, the vector $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\mathbf{w}}$ constitutes the projection of $\mathbf{y}$ onto this subspace). We will see below that the same regression model also arises as a maximum likelihood solution under suitable probabilistic assumptions. Before doing so, we discuss two variants of the above model that are widely used in practice.

3 Ridge Regression and Lasso

We saw above that the simple least squares regression model requires $\mathbf{X}^\top\mathbf{X}$ to be non-singular; indeed, when $\mathbf{X}^\top\mathbf{X}$ is close to being singular (which is the case if two or more columns of $\mathbf{X}$ are nearly co-linear), then

$\widehat{\mathbf{w}}$ can contain large values that lead to over-fitting the training data. To prevent this, one often adds a **penalty** term or a **regularizer** to the objective in Eq. (1) that penalizes large values in $\mathbf{w}$ (such methods are also referred to as **parameter shrinkage** methods in statistics).

A widely used regularizer is the L_2 **regularizer** $\|\mathbf{w}\|_2^2 = \sum_{j=1}^d w_j^2$, leading to the following:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{m}\sum_{i=1}^m \left(\mathbf{w}^\top \mathbf{x}_i - y_i\right)^2 + \lambda\|\mathbf{w}\|_2^2\,, \tag{2}$$

where $\lambda > 0$ is a suitable regularization parameter that determines the trade-off between the two terms. Setting the gradient of the above objective to zero again yields a closed-form solution for $\mathbf{w}$:

$$\widehat{\mathbf{w}} = \left(\mathbf{X}^\top\mathbf{X} + \lambda m \mathbf{I}_d\right)^{-1}\mathbf{X}^\top\mathbf{y}\,,$$

where $\mathbf{I}_d$ denotes the $d \times d$ identity matrix; note that the matrix $\left(\mathbf{X}^\top\mathbf{X} + \lambda m\mathbf{I}_d\right)$ is non-singular. The resulting regression model,

$$f_S(\mathbf{x}) = \widehat{\mathbf{w}}^\top\mathbf{x}\,,$$

is known as **ridge regression** and is widely used in practice.[1]

Another regularizer that is frequently used is the L_1 **regularizer** $\|\mathbf{w}\|_1 = \sum_{j=1}^d |w_j|$, which leads to

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{m}\sum_{i=1}^m \left(\mathbf{w}^\top \mathbf{x}_i - y_i\right)^2 + \lambda\|\mathbf{w}\|_1\,, \tag{3}$$

where $\lambda > 0$ is again a suitable regularization parameter. This can be formulated as a quadratic programming problem which can be solved using numerical optimization methods. For large enough λ, the solution $\widehat{\mathbf{w}}$ turns out to be **sparse**, in the sense that many of the parameter values in $\widehat{\mathbf{w}}$ are equal to zero, so that the resulting regression model depends on only a small number of features. The L_1-regularized least squares regression model is known as **lasso** and is also widely used, especially in high-dimensional problems where d is large and dependence on a small number of features is desirable.

For both L_1 and L_2 regularizers, the regularization parameter λ determines the extent of the penalty for large values in the parameter vector $\mathbf{w}$. In practice, one generally selects λ heuristically from some finite range using a **validation set** (which involves holding out part of the training data for validation, training on the remaining data with different values of λ, and selecting the one that gives highest performance on the validation data) or **cross-validation** (which involves dividing the training sample into some k subsamples/folds, holding out one of these folds at a time and training on the remaining $k-1$ folds with different values of λ, testing performance on the held-out fold, and repeating this procedure for all k folds; the value of λ that gives the highest average performance over the k folds is then selected[2]). In recent years, algorithms for certain models (including lasso) have been developed that can efficiently compute the entire path of solutions for all values of λ. Below we will also see a Bayesian interpretation of these regularizers; this gives another approach to selecting λ.

4 Probabilistic View

We will now make a specific assumption on the conditional distribution of Y given $\mathbf{x}$, and will see that estimating the parameters of that distribution from the training sample using maximum likelihood estimation and using the conditional expectation associated with the estimated distribution as our regression model

[1] The same regularizer is also widely used in logistic regression, leading to **L_2-regularized logistic regression**.

[2] An extreme case of cross-validation with $k = m$ leads to what is called **leave-one-out validation**.

will recover the linear least squares regression model described above. We will also see that under the same probabilistic assumption, maximum a posteriori (MAP) estimation of the parameters under suitable priors will yield ridge regression and lasso.

Specifically, assume that given $\mathbf{x} \in \mathcal{X}$, a label Y is generated randomly as follows:

$$Y = \mathbf{w}^\top \mathbf{x} + \epsilon \,,$$

where $\mathbf{w} \in \mathbb{R}^d$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is some normally distributed noise with variance $\sigma^2 > 0$. In other words, we have

$$Y | X = \mathbf{x} \;\sim\; \mathcal{N}(\mathbf{w}^\top \mathbf{x}, \sigma^2) \,,$$

so that the conditional density of Y given $\mathbf{x}$ can be written as

$$p(y \,|\, \mathbf{x}; \mathbf{w}, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{(y - \mathbf{w}^\top \mathbf{x})^2}{2\sigma^2} \Big) \,.$$

Clearly, in this case, the optimal regression model (under squared error) is given by

$$f^*(\mathbf{x}) = \mathbf{E}[Y | X = \mathbf{x}] = \mathbf{w}^\top \mathbf{x} \,.$$

In practice, the parameters $\mathbf{w}, \sigma$ are unknown and must be estimated from the training sample $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$, which is assumed to contain examples drawn i.i.d. from the same distribution. Let us first proceed with maximum likelihood estimation.

Maximum likelihood estimation. We can write the conditional likelihood of $\mathbf{w}, \sigma$ as

$$\mathcal{L}(\mathbf{w}, \sigma) \;=\; p\big(y_1, \ldots, y_m \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_m; \mathbf{w}, \sigma\big) \;=\; \prod_{i=1}^m p(y_i \,|\, \mathbf{x}_i; \mathbf{w}, \sigma) \;=\; \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{2\sigma^2} \Big) \,.$$

The log-likelihood becomes

$$\ln \mathcal{L}(\mathbf{w}, \sigma) \;=\; -\frac{m}{2} \ln(2\pi) - m \ln \sigma - \sum_{i=1}^m \frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{2\sigma^2} \,.$$

Clearly, maximizing the above log-likelihood w.r.t. $\mathbf{w}$ is equivalent to simply minimizing the empirical squared error on S, yielding the same solution as above:

$$\widehat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \,.$$

This yields the same linear least squares regression model as above:

$$f_S(\mathbf{x}) = \widehat{\mathbf{w}}^\top \mathbf{x} \,.$$

The variance parameter σ does not play a role in the regression model, but can be useful in determining the uncertainty in the model's prediction at any point. It can be estimated by maximizing the log-likelihood above w.r.t. σ, which gives

$$\widehat{\sigma}^2 = \frac{1}{m} \sum_{i=1}^m \big(y_i - \widehat{\mathbf{w}}^\top \mathbf{x}_i\big)^2 \,.$$

Maximum a posteriori (MAP) estimation. Continuing with the normal (Gaussian) noise model above, we can estimate $\mathbf{w}$ using maximum a posteriori (MAP) estimation under a suitable prior rather than using maximum likelihood estimation. For example, let us assume a zero-mean, isotropic normal prior on $\mathbf{w}$. In other words, denoting by W the random variable corresponding to $\mathbf{w}$, we have

$$W \sim \mathcal{N}\big(\mathbf{0}, \sigma_0^2 \, \mathbf{I}_d\big) \,,$$

where $\mathbf{I}_d$ denotes the $d \times d$ identity matrix; this is equivalent to assuming that the prior selects each component of W independently from a $\mathcal{N}(0, \sigma_0^2)$ distribution. The prior density can be written as

$$p(\mathbf{w}) = \frac{1}{(2\pi)^{d/2}\sigma_0^d} \exp\Big(- \frac{1}{2\sigma_0^2}\|\mathbf{w}\|_2^2 \Big) .$$

Assuming for simplicity that the noise variance parameter σ is known, the posterior density of $\mathbf{w}$ given the data S then takes the form

$$p(\mathbf{w}|S) \;\propto\; \exp\Big(- \frac{1}{2\sigma_0^2}\|\mathbf{w}\|_2^2 \Big) \cdot \prod_{i=1}^{m} \exp\Big(- \frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{2\sigma^2} \Big) ,$$

giving

$$\ln p(\mathbf{w}|S) = -\frac{1}{2\sigma_0^2}\|\mathbf{w}\|_2^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{m} \big(y_i - \mathbf{w}^\top \mathbf{x}_i\big)^2 + \text{const} .$$

The MAP estimate of $\mathbf{w}$ is obtained by maximizing this w.r.t. $\mathbf{w}$; clearly, this is equivalent to solving the following L_2-regularized least squares regression problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \; \frac{1}{m}\sum_{i=1}^{m} \big(\mathbf{w}^\top \mathbf{x}_i - y_i\big)^2 + \frac{\sigma^2}{m\sigma_0^2}\|\mathbf{w}\|_2^2 .$$

This provides an alternative view of ridge regression, and suggests that where it is suitable to assume the above conditional distribution with noise variance σ^2 and an isotropic normal prior on $\mathbf{w}$ with variance σ_0^2, an appropriate choice for the regularization parameter is given by $\lambda = \frac{\sigma^2}{m\sigma_0^2}$.

If instead of an isotropic normal prior we assume an isotropic Laplace prior with density

$$p(\mathbf{w}) = \Big(\frac{\lambda_0}{2}\Big)^d \exp\Big(- \lambda_0 \|\mathbf{w}\|_1 \Big) ,$$

then the posterior density becomes

$$p(\mathbf{w}|S) \;\propto\; \exp\Big(- \lambda_0\|\mathbf{w}\|_1 \Big) \cdot \prod_{i=1}^{m} \exp\Big(- \frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{2\sigma^2} \Big) ,$$

with

$$\ln p(\mathbf{w}|S) = -\lambda_0\|\mathbf{w}\|_1 - \frac{1}{2\sigma^2}\sum_{i=1}^{m} \big(y_i - \mathbf{w}^\top \mathbf{x}_i\big)^2 + \text{const} .$$

In this case, finding the MAP estimate of $\mathbf{w}$ is equivalent to solving the following L_1-regularized least squares regression problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \; \frac{1}{m}\sum_{i=1}^{m} \big(\mathbf{w}^\top \mathbf{x}_i - y_i\big)^2 + \frac{2\sigma^2\lambda_0}{m}\|\mathbf{w}\|_1 .$$

Again, this provides an alternative view of lasso, and suggests that where it is suitable to assume the above conditional distribution with noise variance σ^2 and an isotropic Laplace prior on $\mathbf{w}$ with parameter λ_0, an appropriate choice for the regularization parameter is given by $\lambda = \frac{2\sigma^2\lambda_0}{m}$.

Exercise. Show that for any $f : \mathcal{X} \to \mathbb{R}$, the **squared-error regret** of f, i.e. the difference of its squared error from the optimal, is equal to the expected squared difference between $f(X)$ and $\mathbf{E}[Y|X]$:

$$\text{er}_D^{\text{sq}}[f] - \text{er}_D^{\text{sq},*} \;=\; \mathbf{E}_X\big[\big(f(X) - \mathbf{E}[Y|X]\big)^2\big] .$$