kernel ridge regression

Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland

*Corresponding author: alex.jung@aalto.fi

Abstract

Kernel ridge regression is a kernel method for regression that applies ridge regression to transformed feature vectors constructed from a kernel, and thereby learns a nonlinear hypothesis in closed form. It performs regularized empirical risk minimization (RERM) over the reproducing kernel Hilbert space (RKHS) of the kernel, combining the squared error loss with a squared-norm penalty term. By the representer theorem, the learned hypothesis is a weighted sum of kernel evaluations at the feature vectors in the training set, and the coefficients are obtained in closed form by solving a single system of linear equations. Like ridge regression, the method equals empirical risk minimization (ERM) on an augmented training set: the transformed feature vectors are perturbed by realizations of a zero-mean Gaussian process (GP) whose covariance function is the kernel scaled by the regularization parameter. For the kernel that is a modified inner product of the feature space, kernel ridge regression is ridge regression with a matching penalty term, and the perturbed copies of a data point fill an ellipse whose shape the kernel sets, the circle of ridge regression for the standard inner product.

Definition

Consider a regression problem where the goal is to learn a hypothesis for predicting the numeric label of a data point based on its feature vector, but where the relation between features and label is nonlinear. An example is the prediction of the temperature over the course of a day from the time of the day, a relation that rises and falls over the day. Kernel ridge regression solves such regression problems by applying ridge regression to transformed feature vectors constructed from a kernel, and thereby learns a nonlinear hypothesis in closed form (Hastie et al., 2009, Sect. 5.8.2; Schölkopf and Smola, 2002). Fig. 1 shows a training set with a periodic relation between feature and label. With the Gaussian kernel $k\big(x,x'\big) = \exp\big(-(x-x')^2/(2\sigma^2)\big)$ of bandwidth $\sigma = 0.7$, the learned hypothesis follows the data points smoothly, with a training mean squared error (MSE) of 0.013, while ridge regression on the raw feature can only fit a straight line, with a training MSE of 0.36.

Formally, kernel ridge regression learns a hypothesis from the reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ of a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by minimizing the penalized average squared error loss on a training set $\mathcal{D}^{(\text{train})} = \big\{ \big(\mathbf{x}^{(r)}, y^{(r)}\big) \big\}_{r=1}^{m}$

with numeric labels $y^{(r)} \in \mathbb{R}$,

$$\widehat{h} = \operatorname*{arg\,min}_{h \in \mathcal{H}_k} \frac{1}{m} \sum_{r=1}^{m} \big(y^{(r)} - h\big(\mathbf{x}^{(r)}\big)\big)^2 + \alpha \|h\|_{\mathcal{H}_k}^2. \tag{1}$$

The penalty term is the scaled squared RKHS norm $\alpha\|h\|_{\mathcal{H}_k}^2$ with a regularization parameter $\alpha > 0$. The resulting method is an instance of regularized empirical risk minimization (RERM), as is ridge regression, whose penalty term is the scaled squared Euclidean norm $\alpha\|\mathbf{w}\|_2^2$ of the model parameters (Hastie et al., 2009, Sect. 5.8.2). The average squared error loss is a convex function of h, and the squared norm is strictly convex. The kernel ridge regression problem (1) therefore has a unique minimizer $\widehat{h}$ for any training set, as does ridge regression.

By the representer theorem (see kernel method), the minimizer of the kernel ridge regression problem (1) has the form $\widehat{h} = \sum_{r=1}^{m} \beta_r k\big(\mathbf{x}^{(r)}, \cdot\big)$ with a coefficient vector $\boldsymbol{\beta} = \big(\beta_1, \ldots, \beta_m\big)^\top$. Inserting this form into the kernel ridge regression problem (1) yields the finite-dimensional optimization problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^m} \frac{1}{m} \|\mathbf{y} - \mathbf{K}\boldsymbol{\beta}\|_2^2 + \alpha\, \boldsymbol{\beta}^\top \mathbf{K} \boldsymbol{\beta},$$

with the label vector $\mathbf{y} = \big(y^{(1)}, \ldots, y^{(m)}\big)^\top$ and the Gram matrix $\mathbf{K}$ with entries $K_{r,r'} = k\big(\mathbf{x}^{(r)}, \mathbf{x}^{(r')}\big)$ (see kernel). By the zero-gradient condition, a coefficient vector solving this optimization problem satisfies the system of linear equations

$$\mathbf{K}\big((\mathbf{K} + \alpha m \mathbf{I})\boldsymbol{\beta} - \mathbf{y}\big) = \mathbf{0}.$$

Since $\mathbf{K}$ is positive semi-definite (psd) and $\alpha > 0$, the matrix $\mathbf{K} + \alpha m \mathbf{I}$ is invertible, and a minimizer is obtained in closed form as

$$\widehat{\boldsymbol{\beta}} = \big(\mathbf{K} + \alpha m \mathbf{I}\big)^{-1} \mathbf{y}. \tag{2}$$

The prediction for a data point with feature vector $\mathbf{x}$ is

$$\widehat{h}(\mathbf{x}) = \sum_{r=1}^{m} \widehat{\beta}_r k\big(\mathbf{x}^{(r)}, \mathbf{x}\big).$$

Training is a one-shot computation: kernel ridge regression requires no iterative optimization method (and hence no fixed-point interpretation), only the unique solution of the $m \times m$ system of linear equations with the invertible matrix $\mathbf{K} + \alpha m \mathbf{I}$, which is the closed form (2).

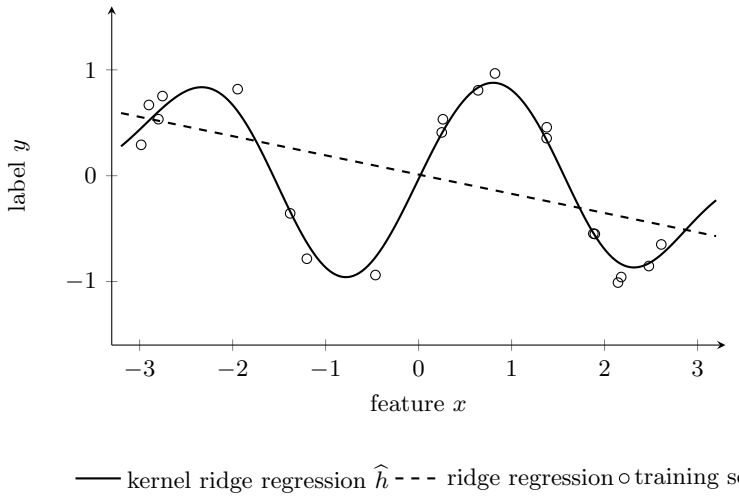

Figure 1: Kernel ridge regression with a Gaussian kernel ($\sigma = 0.7$, $\alpha = 10^{-2}$) on a training set of $m = 20$ noisy data points (circles) generated from a periodic function. The learned hypothesis $\widehat{h}$ (solid) follows the nonlinear relation; ridge regression on the raw feature (dashed) can only fit a straight line. Its training MSE is more than ten times larger than that of kernel ridge regression. Data generated by `pythondemos/kernelridgeregression.py`

For the linear kernel $k\big(\mathbf{x}, \mathbf{x}'\big) = \mathbf{x}^\top \mathbf{x}'$ on $\mathcal{X} = \mathbb{R}^d$, kernel ridge regression produces the same predictions as ridge regression applied to the raw feature vectors (Rasmussen and Williams, 2006, Sect. 2.8). Kernel ridge regression is also closely related to Gaussian process (GP) regression: the prediction $\widehat{h}(\mathbf{x})$ coincides with the posterior mean of a GP with covariance function k and noise variance αm (Kanagawa et al., 2018, Prop. 3.6).

Like ridge regression, kernel ridge regression has an interpretation as a form of data augmentation: training with noise added to the feature vectors is equivalent to adding a penalty term (Bishop, 1995), and replacing each data point by a probability distribution centered at it is the vicinal form of empirical risk minimization (ERM) (Chapelle et al., 2001), whose effect on kernel methods is a regularization of the learned hypothesis (Dao et al., 2019). Consider data points with two features, the feature space $\mathcal{X} = \mathbb{R}^2$, and the kernel

$$k\big(\mathbf{x}, \mathbf{x}'\big) := \mathbf{x}^\top \mathbf{C}^{-1} \mathbf{x}', \tag{3}$$

with a positive definite (pd) matrix $\mathbf{C} \in \mathbb{R}^{2\times 2}$. This kernel is an inner product of $\mathbb{R}^2$ with a modified metric: the squared norm $k\big(\mathbf{x}, \mathbf{x}\big) = \mathbf{x}^\top \mathbf{C}^{-1} \mathbf{x}$ scales the two features differently and couples them, and its unit ball is an ellipse instead of the circle of the Euclidean norm. The RKHS $\mathcal{H}_k$ of this kernel is $\mathbb{R}^2$ with this inner product. The transformed feature vector $k\big(\mathbf{x}, \cdot\big)$ is the linear function $\mathbf{x}' \mapsto \mathbf{x}^\top \mathbf{C}^{-1} \mathbf{x}'$, every hypothesis in $\mathcal{H}_k$ is a linear model $h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ with

model parameters $\mathbf{w} \in \mathbb{R}^2$, and its RKHS norm is $\left\|h^{(\mathbf{w})}\right\|_{\mathcal{H}_k}^2 = \mathbf{w}^\top \mathbf{C}\mathbf{w}$. The kernel ridge regression problem (1) therefore reads

$$\widehat{\mathbf{w}} = \operatorname*{arg\,min}_{\mathbf{w}\in\mathbb{R}^2} \frac{1}{m}\sum_{r=1}^{m} \left(y^{(r)} - \mathbf{w}^\top \mathbf{x}^{(r)}\right)^2 + \alpha\, \mathbf{w}^\top \mathbf{C}\mathbf{w}, \tag{4}$$

which is ridge regression with the penalty term $\alpha\|\mathbf{w}\|_2^2$ replaced by $\alpha\, \mathbf{w}^\top \mathbf{C}\mathbf{w}$. Its unique minimizer is $\widehat{\mathbf{w}} = \left(\mathbf{X}^\top\mathbf{X} + \alpha m \mathbf{C}\right)^{-1}\mathbf{X}^\top\mathbf{y}$ with the feature matrix $\mathbf{X} = \left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\right)^\top$, and it coincides with $\mathbf{C}^{-1}\mathbf{X}^\top\widehat{\boldsymbol{\beta}}$ for the coefficients (2) with the Gram matrix $\mathbf{K} = \mathbf{X}\mathbf{C}^{-1}\mathbf{X}^\top$. For $\mathbf{C} = \mathbf{I}$, the kernel (3) is the linear kernel and the penalized problem (4) is ridge regression itself.

Adding the penalty term in the penalized problem (4) is equivalent to ERM on an augmented training set, i.e., to computing the average squared error loss of $h^{(\mathbf{w})}$ on it. This augmented training set is obtained from the original training set. For each data point $\left(\mathbf{x}^{(r)}, y^{(r)}\right)$, $r = 1, \ldots, m$, the feature vector $\mathbf{x}^{(r)}$ is replaced by realizations of infinitely many independent and identically distributed (i.i.d.) random variables (RVs) $\mathbf{x}^{(r)} + \boldsymbol{\varepsilon}^{(r)}$ with Gaussian random variables (Gaussian RVs) $\boldsymbol{\varepsilon}^{(r)} \sim \mathcal{N}\left(\mathbf{0}, \alpha\mathbf{C}\right)$. The label $y^{(r)}$ is left unchanged. The penalty term is exactly the amount by which the average squared error loss of $h^{(\mathbf{w})}$ over the perturbed copies exceeds its squared error loss on the original data point. Since $\mathbb{E}\{\boldsymbol{\varepsilon}^{(r)}\} = \mathbf{0}$, the increase is

$$\mathbb{E}\left\{\left(\mathbf{w}^\top\boldsymbol{\varepsilon}^{(r)}\right)^2\right\} = \mathbf{w}^\top\mathbb{E}\left\{\boldsymbol{\varepsilon}^{(r)}\left(\boldsymbol{\varepsilon}^{(r)}\right)^\top\right\}\mathbf{w} = \alpha\, \mathbf{w}^\top\mathbf{C}\mathbf{w}.$$

The equivalence is exact for the linear model $h^{(\mathbf{w})}$, while for a nonlinear hypothesis it holds up to terms of higher order in the noise amplitude (Bishop, 1995). As in ridge regression, this increase of the average loss under data augmentation is an estimate for the generalization gap. Accordingly, linear regression on a large augmented training set recovers the predictions of kernel ridge regression.

The kernel determines the shape of these perturbations (see Fig. 2). The covariance matrix $\alpha\mathbf{C}$ of $\boldsymbol{\varepsilon}^{(r)}$ is the scaled inverse of the matrix in the kernel (3), so the one-standard-deviation contour of the perturbed copies,

$$\left(\mathbf{x} - \mathbf{x}^{(r)}\right)^\top \mathbf{C}^{-1}\left(\mathbf{x} - \mathbf{x}^{(r)}\right) = \alpha,$$

is the set of feature vectors $\mathbf{x}$ with $k\left(\mathbf{x} - \mathbf{x}^{(r)}, \mathbf{x} - \mathbf{x}^{(r)}\right) = \alpha$, i.e., the ball of radius $\sqrt{\alpha}$ around $\mathbf{x}^{(r)}$ in the norm induced by the kernel. For $\mathbf{C} = \mathbf{I}$, this ball is a circle of radius $\sqrt{\alpha}$: the perturbations in the data augmentation interpretation of ridge regression have covariance matrix $\alpha\mathbf{I}$ and no preferred direction. For a general $\mathbf{C}$ it is an ellipse, so the ellipses drawn around the data points display the kernel: the size of a perturbation is measured by its kernel norm, and the penalty term $\alpha\, \mathbf{w}^\top\mathbf{C}\mathbf{w}$ is the average squared error loss that perturbations of kernel norm $\sqrt{\alpha}$ cause. The ellipses also display the matrix $\mathbf{C}$ itself. Its eigenvalue decomposition (EVD) $\mathbf{C} = \sum_{j=1}^{2} \lambda_j \mathbf{u}^{(j)}\left(\mathbf{u}^{(j)}\right)^\top$, with

eigenvalues $\lambda_1 \geq \lambda_2 > 0$ and orthonormal eigenvectors $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}$, turns the contour into

$$\sum_{j=1}^{2} \frac{\left(\left(\mathbf{u}^{(j)}\right)^\top \left(\mathbf{x} - \mathbf{x}^{(r)}\right)\right)^2}{\alpha \lambda_j} = 1,$$

so the principal axes of every ellipse point along the eigenvectors of $\mathbf{C}$ and the semi-axes have lengths $\sqrt{\alpha\lambda_1}$ and $\sqrt{\alpha\lambda_2}$: the perturbations spread most along the eigenvector with the largest eigenvalue, and the penalty term charges a displacement along $\mathbf{u}^{(j)}$ at the rate λ_j.

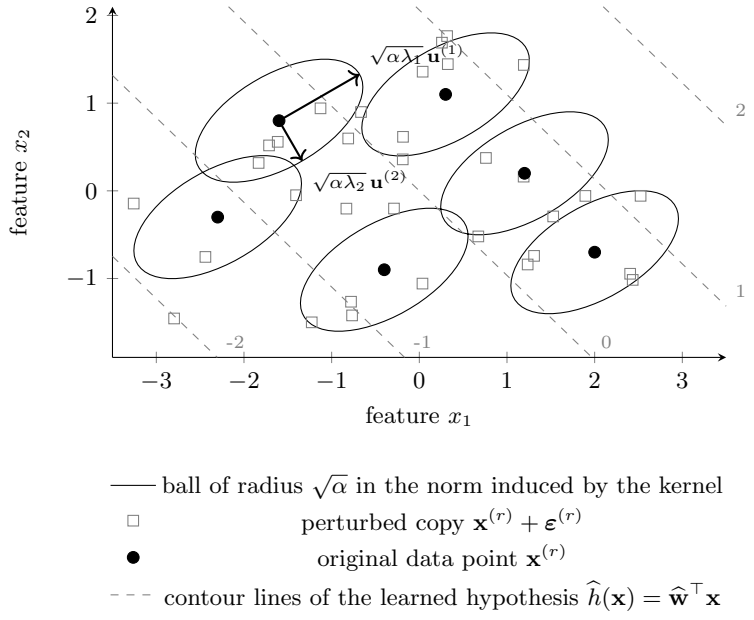

Figure 2: The data augmentation interpretation of kernel ridge regression with the kernel (3), for a training set of six data points with two features, $\alpha = 1/2$, and a matrix $\mathbf{C}$ with eigenvalues $\lambda_1 = 1.5^2$, $\lambda_2 = 0.75^2$ and eigenvectors rotated by 30° against the feature axes. Each original data point (filled circle) is replaced by copies (open squares) whose feature vectors are perturbed by Gaussian RVs with covariance matrix $\alpha\mathbf{C}$, keeping the label. The ellipse around each data point is the one-standard-deviation contour of its copies, the ball of radius $\sqrt{\alpha}$ in the norm induced by the kernel; for $\mathbf{C} = \mathbf{I}$ it is the circle of radius $\sqrt{\alpha}$ of the isotropic perturbations of ridge regression. The arrows on one ellipse are its principal axes $\sqrt{\alpha\lambda_j}\,\mathbf{u}^{(j)}$, the eigenvectors of $\mathbf{C}$ scaled by the square roots of α times its eigenvalues. The dashed lines are contour lines of the hypothesis $\widehat{h}(\mathbf{x}) = \widehat{\mathbf{w}}^\top \mathbf{x}$ learned from the six data points by the penalized problem (4), labelled with its value: a copy displaced along a contour line keeps its prediction, and a copy displaced across the contour lines changes it in proportion to the displacement. Data generated by `pythondemos/kernelridgeregression.py`

For the kernel (3), the perturbation acts linearly on the transformed feature vectors, $k\big(\mathbf{x}^{(r)} + \boldsymbol{\varepsilon}^{(r)}, \cdot\big) = k\big(\mathbf{x}^{(r)}, \cdot\big) + k\big(\boldsymbol{\varepsilon}^{(r)}, \cdot\big)$, and the random function $k\big(\boldsymbol{\varepsilon}^{(r)}, \cdot\big)$ is a zero-mean GP whose covariance function is the kernel scaled by the regularization parameter,

$$\mathbb{E}\big\{k\big(\boldsymbol{\varepsilon}^{(r)}, \mathbf{x}\big)\, k\big(\boldsymbol{\varepsilon}^{(r)}, \mathbf{x}'\big)\big\} = \mathbf{x}^\top \mathbf{C}^{-1} \mathbb{E}\big\{\boldsymbol{\varepsilon}^{(r)} \big(\boldsymbol{\varepsilon}^{(r)}\big)^\top\big\} \mathbf{C}^{-1} \mathbf{x}' = \alpha\, k\big(\mathbf{x}, \mathbf{x}'\big). \quad (5)$$

For a general kernel, the raw feature vectors cannot be perturbed in this way, but the perturbation of the transformed feature vector $k\big(\mathbf{x}^{(r)}, \cdot\big)$ by a zero-mean GP with covariance function αk remains defined (Kanagawa et al., 2018). Up to the scaling by α, this is the GP whose posterior mean kernel ridge regression computes.

Cross References

kernel method, ridge regression, kernel, RKHS, RERM, squared error loss, regression, GP, data augmentation, feature transformation, penalty term

References

1. Bishop (1995). Training with Noise is Equivalent to Tikhonov Regularization. Neural Computation, 7(1), pp. 108–116.

2. Chapelle et al. (2001). Vicinal Risk Minimization. In: Advances in Neural Information Processing Systems 13 (NIPS 2000), pp. 416–422. MIT Press.

3. Dao et al. (2019). A Kernel Theory of Modern Data Augmentation. In: Proceedings of the 36th International Conference on Machine Learning (ICML), pp. 1528–1537. PMLR.

4. Schölkopf and Smola (2002). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA.

5. Rasmussen and Williams (2006). Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA, USA.

6. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.

7. Kanagawa et al. (2018). Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences. URL: https://arxiv.org/abs/1807.02582.