EE269
Signal Processing for Machine Learning

Lecture 13

Instructor : Mert Pilanci

Stanford University

February 27, 2019

Gaussian regression models

- $y \in \mathbb{R}$ continuous label and $x \in \mathbb{R}^d$
- training set $x_1, ...x_n$ and labels $y_1, ..., y_n$

$$p(y|x, w, b) = N(w^T x + b, \sigma^2)$$

- $P(w)$ prior probability of w
- infinitely many classes parametrized by w
- $\max_w \ P(y|x_1, ...x_n) = P(y|x_1, ...x_n, w, b)P(w)$
- independent observations: $= \prod_{i=1}^n P(y_i|x_i, w, b)P(w)$
- Maximum a Posteriori (MAP) estimate w_{MAP}

$$w_{MAP} = \arg\max \prod_{i=1}^{n} P(y_i|x_i, w, b)P(w)$$

$$= \arg\max \sum_{i=1}^{n} \log P(y_i|x_i, w, b) + \log P(w)$$

Gaussian regression models

- $y \in \mathbb{R}$ continuous label and $x \in \mathbb{R}^d$
- training set $x_1, ...x_n$ and labels $y_1, ..., y_n$

$$p(y|x, w, b) = N(w^T x + b, \sigma^2)$$

$$w_{MAP} = \arg\max \sum_{i=1}^{n} \log P(y_i|x_i, w, b) + \log P(w)$$

- Gaussian prior on w : $P(w) = N(0, t^2 I)$

$$w_{MAP} = \arg\max -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w^T x_i - b)^2 - \frac{1}{2t^2} ||w||_2^2$$

Gaussian regression models

- $y \in \mathbb{R}$ continuous label and $x \in \mathbb{R}^d$
- training set $x_1, ...x_n$ and labels $y_1, ..., y_n$

$$p(y|x, w, b) = N(w^T x + b, \sigma^2)$$

$$w_{MAP} = \arg\max_w \sum_{i=1}^{n} \log P(y_i|x_i, w, b) + \log P(w)$$

- Gaussian prior on w : $P(w) = N(0, t^2 I)$

$$w_{MAP} = \arg\min_w \sum_{i=1}^{n} (y_i - w^T x_i - b)^2 + \frac{\sigma^2}{t^2} ||w||_2^2$$

- ℓ_2 regularization (Ridge regression)

Gaussian regression models

- Laplace prior $P(w) \propto e^{-\frac{|w_1|}{t}} ... e^{-\frac{|w_d|}{t}}$

$$w_{MAP} = \arg\min_w \sum_{i=1}^{n} (y_i - w^T x_i - b)^2 + \frac{\sigma^2}{t} ||w||_1$$

- ℓ_1 regularization (Lasso)

ℓ_2 regularization (Ridge regression)

$$\min_{w} \; ||Xw - y||_2^2 + \lambda ||w||_2^2 \tag{1}$$

ℓ_1 regularization (Lasso)

$$\min_{w} \; ||Xw - y||_2^2 + \lambda ||w||_1 \tag{2}$$

Exponential density $e^{-\frac{|w|}{t}}$ vs Gaussian density $e^{-\frac{w^2}{t^2}}$

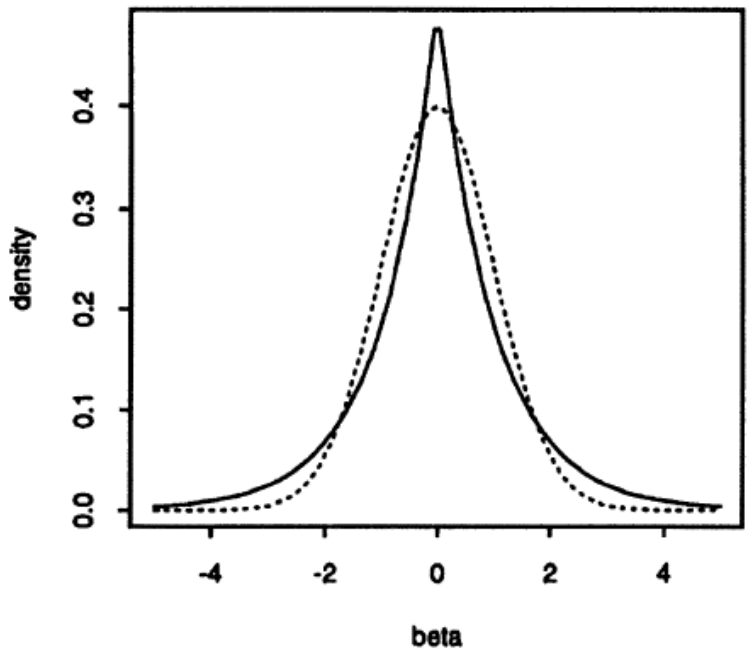


Fig. 7. Double-exponential density (———) and normal density (- - - - -): the former is the implicit prior used by the lasso; the latter by ridge regression

Least Squares Regression and Duality

- in matrix-vector notation (redefine $\lambda \leftarrow n\lambda$)

$$\min_{w} \ \frac{1}{2}||Xw - y||_2^2 + \frac{\lambda}{2}||w||_2^2$$

- equivalent constrained problem

$$\min_{z,w \ : \ Xw=z} \ \frac{1}{2}||z - y||_2^2 + \frac{\lambda}{2}||w||_2^2$$

- Dual problem:

$$\max_{\alpha} -\alpha^T(\frac{1}{2\lambda}XX^T + \frac{1}{2}I)\alpha + \alpha^T y$$

KKT conditions imply $w^* = \frac{1}{\lambda}X^T\alpha^*$

We can solve the dual in closed form $\alpha^* = (\frac{1}{\lambda}XX^T + I)^{-1}y$

Dual Least Squares Regression Problem

Dual problem:

$$\max_{\alpha} -\alpha^T(\frac{1}{2\lambda}XX^T + \frac{1}{2}I)\alpha + \alpha^T y$$

KKT conditions imply $w^* = \frac{1}{\lambda}X^T\alpha^*$

We can solve the dual in closed form $\alpha^* = (\frac{1}{\lambda}XX^T + I)^{-1}y$

- Given test sample x, the prediction is $w^{*T}x = \left(\frac{1}{\lambda}\sum_{i=1}^n x_i\alpha_i^*\right)x = \frac{1}{\lambda}\sum_{i=1}^n \langle x_i, x\rangle\alpha_i^*$
- Kernel map $x \to \Phi(x)$ and kernel matrix $K_{ij} = \kappa(x, y) = \langle\Phi(x), \Phi(x)\rangle$
- Dual solution $\alpha^* = (\frac{1}{\lambda}K + I)^{-1}y$
- Prediction $\hat{f}(x) = \frac{1}{\lambda}\sum_{i=1}^n \kappa(x_i, x)\alpha_i^*$

Kernel Regression Application

- polynomial kernel $\kappa(x, y) = (1 + x^T y)^4$
  prediction $\hat{f}(x) = \frac{1}{\lambda} \sum_{i=1}^{n} \kappa(x_i, x) \alpha_i^*$

Kernel Regression Application

- Gaussian kernel $\kappa(x, y) = e^{-\frac{||x-y||_2^2}{2\sigma^2}}$
  prediction $\hat{f}(x) = \frac{1}{\lambda} \sum_{i=1}^{n} \kappa(x_i, x)\alpha_i^*$

Reproducing Kernel Hilbert Space

- Mercer's Theorem: Any positive definite kernel function can be represented in terms of eigenfunctions

$$\kappa(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(y)$$

- The functions $\phi(x)$ form an orthonormal basis for a function space

$$\mathcal{H}_k = \{f \ : \ f(x) = \sum_{i=1}^{\infty} f_i \phi_i(x), \quad \sum_{i=1}^{\infty} \frac{f_i^2}{\lambda_i} < \infty\}$$

Reproducing Kernel Hilbert Space

- Mercer's Theorem: Any positive definite kernel function can be represented in terms of eigenfunctions

$$\kappa(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(y)$$

- The functions $\phi(x)$ form an orthonormal basis for a function space

$$\mathcal{H}_k = \{f \ : \ f(x) = \sum_{i=1}^{\infty} f_i \phi_i(x), \quad \sum_{i=1}^{\infty} \frac{f_i^2}{\lambda_i} < \infty\}$$

- For two functions $f(x) = \sum_{i=1}^{\infty} f_i \phi_i(x)$ and $g(x) = \sum_{i=1}^{\infty} g_i(x) \phi_i$
- define inner product

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} \frac{f_i g_i}{\lambda_i}$$

- reproducing property: $\langle \kappa(x, \cdot), \kappa(y, \cdot) \rangle_{\mathcal{H}} = \kappa(x, y)$

Representer Theorem in Reproducing Kernel Hilbert Space

$$(*) \qquad \min_{f \in \mathcal{H}_k} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda ||f||^2_{\mathcal{H}_k}$$

- Representer theorem : The optimal solution must have the form $f^*(x) = \sum_{i=1}^{n} \alpha_i \kappa(x, x_i)$

Representer Theorem in Reproducing Kernel Hilbert Space

$$(*) \quad \min_{f \in \mathcal{H}_k} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda ||f||^2_{\mathcal{H}_k}$$

- Representer theorem : The optimal solution must have the form $f^*(x) = \sum_{i=1}^{n} \alpha_i \kappa(x, x_i)$
- Plugging in and applying reproducing property $\langle \kappa(x_i, \cdot), \kappa(x_j, \cdot) \rangle_{\mathcal{H}} = \kappa(x_i, x_j)$, we get

$$(*) = \min_{\alpha} ||K\alpha - y||_2^2 + \lambda \alpha^T K \alpha$$

Representer Theorem in Reproducing Kernel Hilbert Space

$$(*) \quad \min_{f \in \mathcal{H}_k} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda ||f||^2_{\mathcal{H}_k}$$

- Representer theorem : The optimal solution must have the form $f^*(x) = \sum_{i=1}^{n} \alpha_i \kappa(x, x_i)$
- Plugging in and applying reproducing property $\langle \kappa(x_i, \cdot), \kappa(x_j, \cdot) \rangle_{\mathcal{H}} = \kappa(x_i, x_j)$, we get

$$(*) = \min_{\alpha} ||K\alpha - y||_2^2 + \lambda \alpha^T K \alpha$$

- solution $\alpha^* = (K + \lambda I)^{-1} y$
- prediction $\hat{f}(x) = \sum_{i=1}^{n} \alpha_i^* \kappa(x, x_i)$
- same prediction rule obtained with dual ridge regression

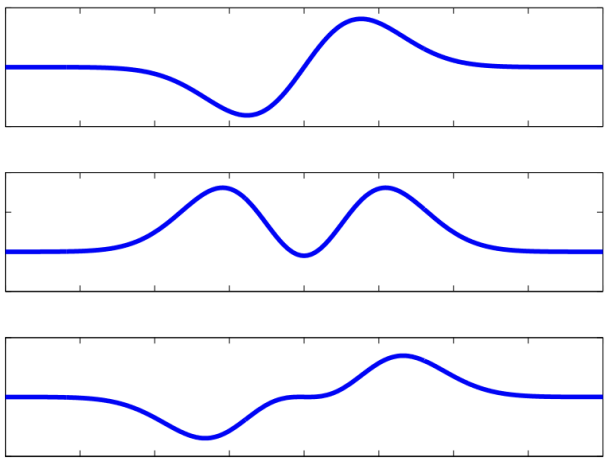
Example: Gaussian Kernel

- Gaussian kernel $\kappa(x, y) = e^{-\frac{||x-y||_2^2}{2\sigma^2}}$

  $\kappa(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(y)$

  $\phi(x) \propto e^{-(c-a)x^2} H_i(x\sqrt{2c})$ and $\lambda_i = b^i$

  a, b, c are functions of σ and H_k is the ith order Hermite polynomial



Example: Gaussian Kernel

- Gaussian kernel $\kappa(x, y) = e^{-\frac{||x-y||_2^2}{2\sigma^2}}$
  $f(x) = \sum_{i=1}^{n} \alpha_i \kappa(x_i, x) = \sum_{i=1}^{n} \sum_{j=1}^{\infty} \lambda_j \phi_j(x_i)\phi_j(x) = \sum_{j=1}^{\infty} f_j \sqrt{\lambda_j}\phi_j(x)$
  where $f_j = \sum_{i=1}^{m} \alpha_i \sqrt{\lambda_j}\phi_j(x_i)$

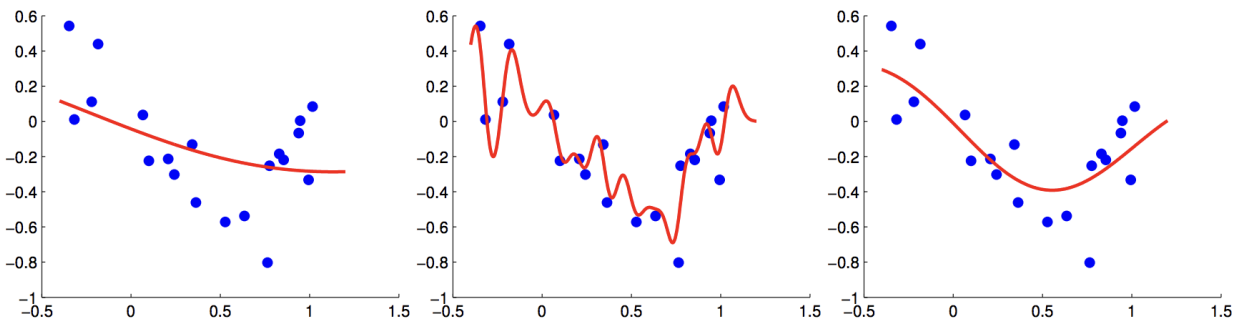
$$\min_{f \in \mathcal{H}_k} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda ||f||_{\mathcal{H}_k}^2$$

Example: Gaussian Kernel

- Gaussian kernel $\kappa(x,y) = e^{-\frac{||x-y||_2^2}{2\sigma^2}}$
  $f(x) = \sum_{i=1}^{n} \alpha_i \kappa(x_i, x) = \sum_{i=1}^{n} \sum_{j=1}^{\infty} \lambda_j \phi_j(x_i)\phi_j(x) = \sum_{j=1}^{\infty} f_j \sqrt{\lambda_j}\phi_j(x)$
  where $f_j = \sum_{i=1}^{m} \alpha_i \sqrt{\lambda_j}\phi_j(x_i)$

$$\min_{f \in \mathcal{H}_k} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda ||f||_{\mathcal{H}_k}^2$$

- For a function $h(x) = \sum_{i=1}^{\infty} h_i \phi_i(x)$
- $||h||_{\mathcal{H}_k}^2 = \langle h, h \rangle_{\mathcal{H}_k} = \sum_{i=1}^{\infty} \frac{h_i^2}{\lambda_i}$ enforces smoothness by penalizing rough eigenfunctions (small λ_i).

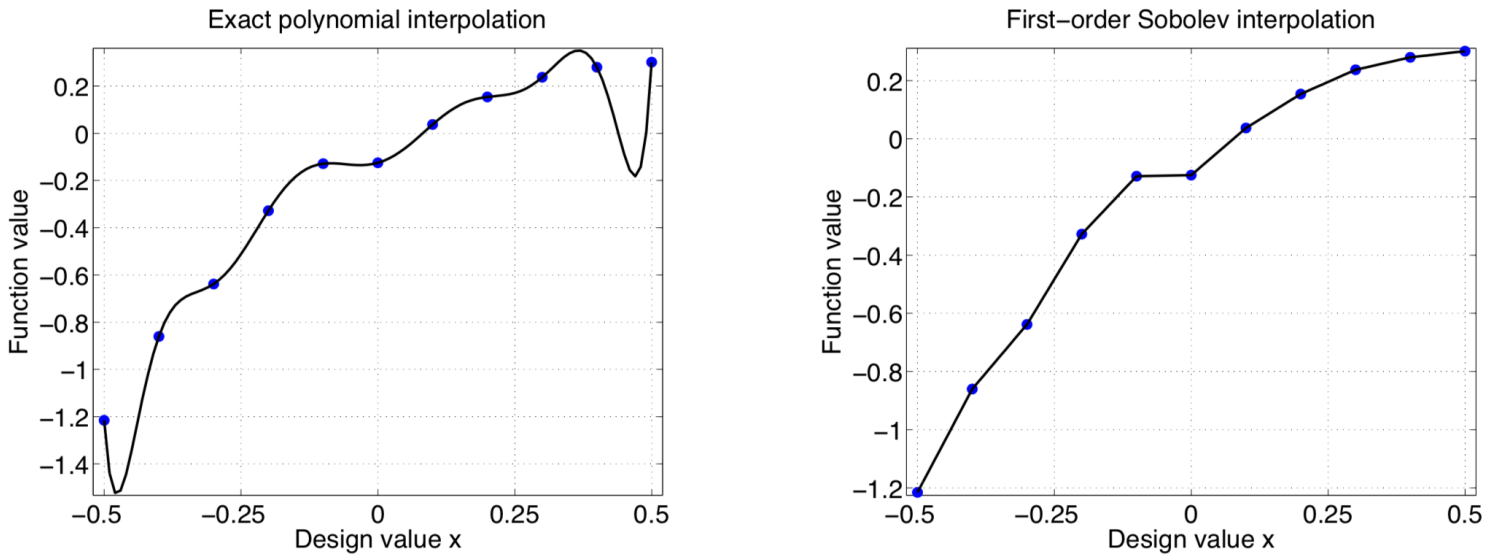


Example: Sobolev Kernel (one dimensional signals)

$$\mathcal{H}_k = \{f \ : [0,1] \to \mathbb{R} \mid f(0) = 0, \text{abs. continuous}, \int_{-\infty}^{\infty} |f'(t)|^2 dt < \infty\}$$

- absolutely continuous $\triangleq$ $f'(t)$ exists almost everywhere,
- $\mathcal{H}_k$ is a Reproducing Kernel Hilbert Space with kernel $\kappa(x, y) = \min(x, y)$

Sobolev Kernel vs Polynomial Kernel



(a) (b)

Figure 12-1. Exact interpolation of $n = 11$ equally sampled function values using RKHS methods. (a) Polynomial kernel $\mathcal{K}(x,\, z) = (1 + x\, z)^{15}$. (b) First-order Sobolev kernel $\mathcal{K}(x,\, z) = 1 + \min\{x, z\}$.

Example: Sinc Kernel (one dimensional signals)

- Paley-Wiener space
- $\kappa(x, y) \triangleq \mathrm{sinc}(\alpha(x - y)) = \frac{\sin(\alpha(x-y))}{\alpha(x-y)}$

Example: Sinc Kernel (one dimensional signals)

- Paley-Wiener space
- $\kappa(x, y) \triangleq \mathrm{sinc}(\alpha(x - y)) = \frac{\sin(\alpha(x-y))}{\alpha(x-y)}$
- $f(t)$: bandlimited functions

Example: Sinc Kernel (one dimensional signals)

- Paley-Wiener space
- $\kappa(x,y) \triangleq \mathrm{sinc}(\alpha(x-y)) = \frac{\sin(\alpha(x-y))}{\alpha(x-y)}$
- $f(t)$: bandlimited functions
- related to Shannon-Whittaker interpolation formula

  uniform samples $x[n] = x(nT)$
- bandlimited interpolation

$$x(t) = \sum_n x[n]\mathrm{sinc}\Big(\frac{t-nT}{T}\Big)$$