Linear Models for Regression

CSci 5525: Machine Learning

Instructor: Paul Schrater

Linear Models

- Linear models over feature representations ϕ_j

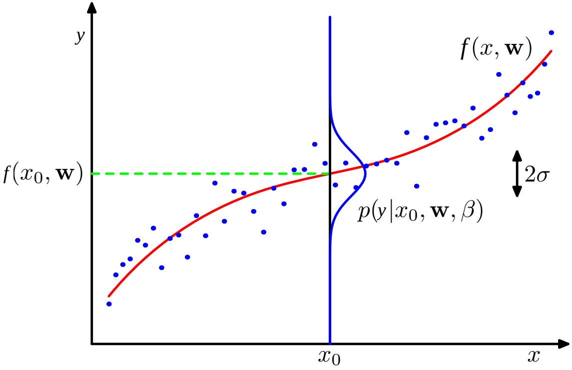
$$f(\mathbf{x}, \mathbf{w}) = \sum_{j=1}^{M} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$$

- Choice of representations: fixed, implicit, learned
- Least squares regression: $y \sim \mathcal{N}(f(\mathbf{x}, \mathbf{w}), \beta^{-1})$

$$p(y|\mathbf{x}, \mathbf{w}, \beta) = \sqrt{\frac{\beta}{2\pi}} \exp\left\{-\frac{\beta}{2}(y - \mathbf{w}^T \phi(\mathbf{x}))^2\right\}$$

- We will often use $\mathbf{x}$ (instead of $\phi(\mathbf{x})$) to denote the feature representation

Conditional Distribution



Maximum Likelihood

- Training set: $(X, \mathbf{y}) = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$
- Assuming statistical independence

$$p(\mathbf{y}|X, \mathbf{w}, \beta) = \prod_{n=1}^{N} p(y_n|\mathbf{x}_n, \mathbf{w}, \beta)$$

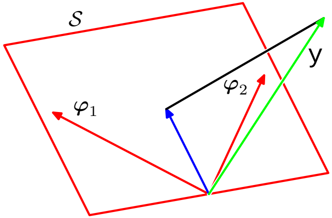
- Goal: Choose $\mathbf{w}$ to maximize the likelihood
- Equivalently, minimize squared loss in terms of $\mathbf{w}$

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} (y_n - \mathbf{w}^T \phi(\mathbf{x}_n))^2$$

- Denoting $\Phi \in \mathbb{R}^{N \times M}$ feature matrix

$$\mathbf{w}^* = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}$$

Geometry of least squares



- $\mathcal{S}$: subspace spanned by basis functions (vectors)
- Least squares: orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}$

Loss Decomposition

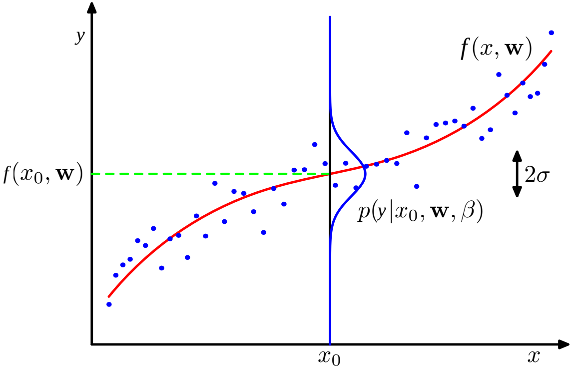
- Total expected loss

$$E_{(\mathbf{x},y)}[\ell(f(\mathbf{x}),y)] = \int\int \ell(f(\mathbf{x}),y)p(\mathbf{x},y)d\mathbf{x}dy$$
$$= \int\int (y - f(\mathbf{x}))^2 p(\mathbf{x},y)d\mathbf{x}dy$$

- Solution $f(\mathbf{x}) = \int yp(y|\mathbf{x}) = E[y|\mathbf{x}]$
- Loss decomposition

$$E_{(\mathbf{x},y)}[\ell(f(\mathbf{x}),y)] = \int (f(\mathbf{x}) - E[y|\mathbf{x}])^2 p(\mathbf{x})d\mathbf{x} + \underbrace{\int\int (E[y|\mathbf{x}] - y)^2 p(\mathbf{x},y)d\mathbf{x}dy}_{\text{distribution variance}}$$

Conditional Distribution



Bias Variance Decomposition

- Let $h(\mathbf{x}) = E[y|\mathbf{x}]$, the best we can do
- For a particular dataset D, we learn $f(\mathbf{x}) = f(\mathbf{x}; D)$
- Taking expectation over all such datasets

$$E_D[(f(\mathbf{x}, D) - h(\mathbf{x}))^2] = \underbrace{(E_D[f(\mathbf{x}; D)] - h(x))^2}_{(\text{bias})^2} + \underbrace{E_D[(f(x; D) - E_D[f(x; D)])^2]}_{\text{variance}}$$

- The overall loss decomposition

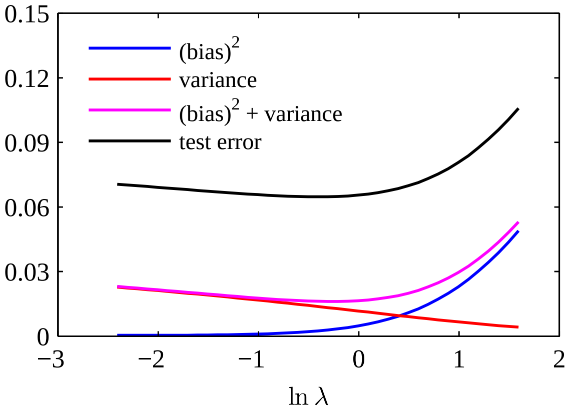
$$E_(\mathbf{x}, y)[\ell(f(\mathbf{x}), y)] = (\text{bias})^2 + \text{variance} + \text{distribution variance}$$

$$(\text{bias})^2 = \int (E_D[f(\mathbf{x}; D)] - h(x))^2 p(\mathbf{x}) d\mathbf{x}$$

$$\text{variance} = \int E_D[(f(\mathbf{x}; D) - E_D[f(\mathbf{x}; D)])^2] p(\mathbf{x}) d\mathbf{x}$$

$$\text{distribution variance} = \int\int (h(\mathbf{x}) - y)^2 p(\mathbf{x}, y) d\mathbf{x} dy$$

Bias Variance Tradeoff



Ridge Regression

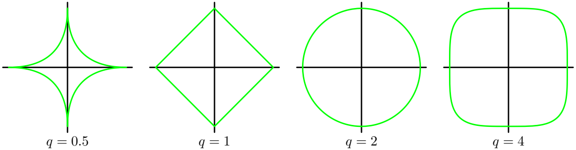
$$\frac{1}{2}\sum_{n=1}^{N}(y_n - \mathbf{w}^T\phi(\mathbf{x}_n))^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

Regularized least squares

- Regularization to control over-fitting

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

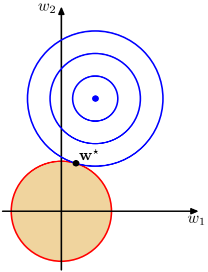
$$\frac{1}{2}\sum_{n=1}^{N}(y_n - \mathbf{w}^T\phi(\mathbf{x}_n))^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$



General classes of regularizers:

$$\frac{1}{2}\sum_{n=1}^{N}(y_n - \mathbf{w}^T\phi(\mathbf{x}_n))^2 + \lambda\|\mathbf{w}\|_{\mathcal{H}}$$

Regularized least squares: Sparse Models



- Regression with L_1 regularization: Lasso

$$\frac{1}{2}\sum_{n=1}^{N}(y_n - \mathbf{w}^T\phi(\mathbf{x}_n))^2 + \frac{\lambda}{2}\|\mathbf{w}\|_1$$

- Regression with "atomic norm" regularization

$$\frac{1}{2}\sum_{n=1}^{N}(y_n - \mathbf{w}^T\phi(\mathbf{x}_n))^2 + \lambda\|\mathbf{w}\|_{\mathcal{A}}$$