Machine Learning EN.601.475/675

DR. PHILIP GRAFF

Linear Regression

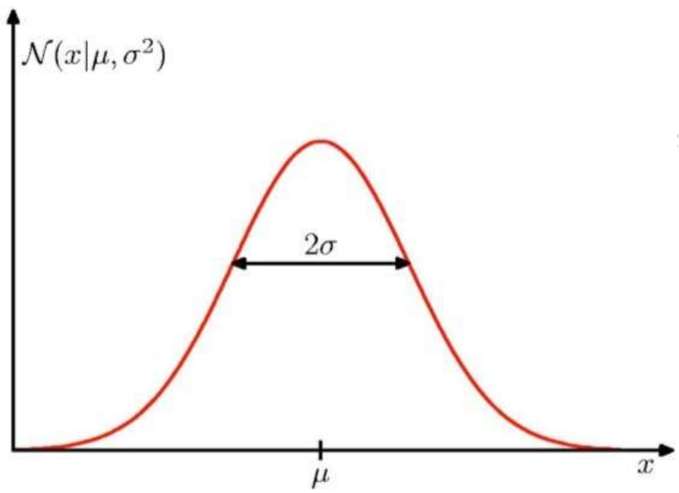
OUR FIRST ALGORITHM!

Noise from a Gaussian Distribution

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$E[x] = \mu$$

$$\mathrm{var}[x] = \sigma^2$$



Noise

Assume the output is perturbed by Gaussian noise

$$y = h_w(x) + \epsilon$$

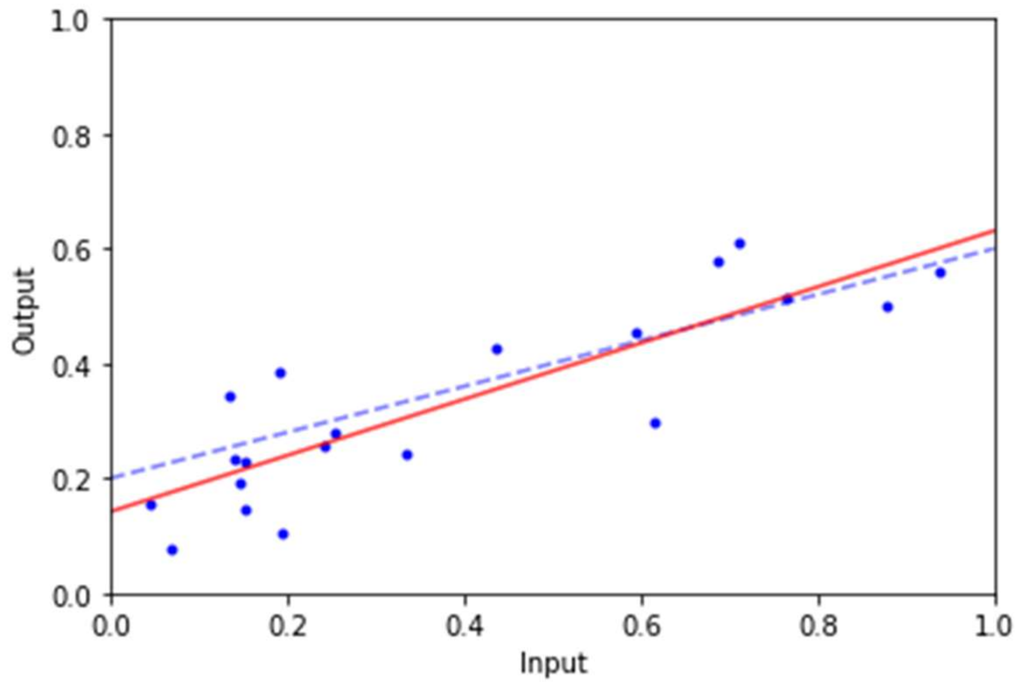
$$\epsilon \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mu = 0, \quad \sigma^2 = 1$$

Not really generated this way

- But we can make this assumption for the sake of modeling

Linear Regression with Noise



Probability of Output

What is the probability of a given predicted output?

- How well does the error match the noise model?

$$
\begin{aligned}
\Pr(y|\mathbf{x}, \mathbf{w}, \sigma^2) &= \mathcal{N}(y, \sigma^2) \\
&= \mathcal{N}(h_{\mathbf{w}}(\mathbf{x}), \sigma^2)
\end{aligned}
$$

Least Squares regression

Fitting: Solve for w given x and y

Function: Linear function + Gaussian noise

- Loss function: squared error
- Assumes output (mostly) a linear combination of input

Data: fit a model to training data, evaluate on held out data

Minimize a function

- What function are we minimizing?

Which Function are we Minimizing?

Which is the best hypothesis?

- Which setting of the parameters **w** is best?

1. Select the hypothesis that best explains the observed data
2. Select the hypothesis that minimizes the error

Explaining the Data

What does it mean to explain the data?

- Maximize the likelihood of the data
- Likelihood = probability of observing the data

Writing the likelihood

- Assume the data is generated from our linear regression model

Maximum Likelihood for Gaussians

$$p(\mathbf{X}|\mu, \sigma^2) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-(x_n - \mu)^2}{2\sigma^2} \right\}$$

$$\ln p(\mathbf{X}|\mu, \sigma^2) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2$$

Maximize with respect to μ: $\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$

Maximize with respect to σ: $\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2$

Maximum Likelihood for Gaussians

$$p(\mathbf{X}|\mu, \mathbf{\Sigma}) = \prod_{n=1}^{N} \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_n - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}_n - \mu)\right\}$$

$$\ln p(\mathbf{X}|\mu, \mathbf{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \mu)^T \mathbf{\Sigma}^{-\mathbf{1}}(\mathbf{x}_n - \mu)$$

Maximize with respect to μ: $$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$$

Maximize with respect to Σ: $$\mathbf{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \mu_{\mathrm{ML}})(\mathbf{x}_n - \mu_{\mathrm{ML}})^T$$

Maximum Likelihood for Regression

Let $\beta^{-1} = \sigma^2$ and $t = y(\mathbf{x}, \mathbf{w}) + \epsilon$

- *t* is the target value, ϵ is the Gaussian noise

$y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$ where ϕ_j are the basis feature functions and $\phi_0(\mathbf{x}) = 1$

Likelihood: $$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n | \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$$

$$\boxed{\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w}) \quad \text{with} \quad E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left(t_n - \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \right)^2}$$

Maximum Likelihood

Take the gradient of the log-likelihood and set to zero to maximize

$$\nabla \ln p(\mathbf{t}|\mathbf{w}, \beta) = \beta \sum_{n=1}^{N} \left(t_n - \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n)\right) \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}$$

$$\mathbf{w}_{\mathrm{ML}} = \left(\mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^{\mathrm{T}} \mathbf{t}$$

$$\mathbf{\Phi}^{\dagger} = \left(\mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^{\mathrm{T}}$$ is the Moore-Penrose pseudo-inverse

Sources of Error

Noise error

- An example has an incorrect or inconsistent label
- Our data representation fails to encode necessary information

Model error

- Hypothesis class is deficient

Parameter estimation error

- The model parameters are wrong

Search error

- We made a mistake in scoring a prediction
- Common in tasks with complex output

Bias?

Gaussians: maximum likelihood estimate is always biased

- This is OK if we have infinite data
- But… we *never* have infinite data!

Over-fitting: avoid it by favoring certain solutions

Regularization

- Add a term to the objective function to favor certain solutions, but what?
  - Occam's Razor: simpler is better
  - Favor small weights → our parameters should be small

Regularized Least Squares

Tradeoff between low error and small weights

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

$$E_W(\mathbf{w}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n))^2 + \lambda\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

Maximum Likelihood Solution

Again, take gradient, set to zero, and solve for **w**

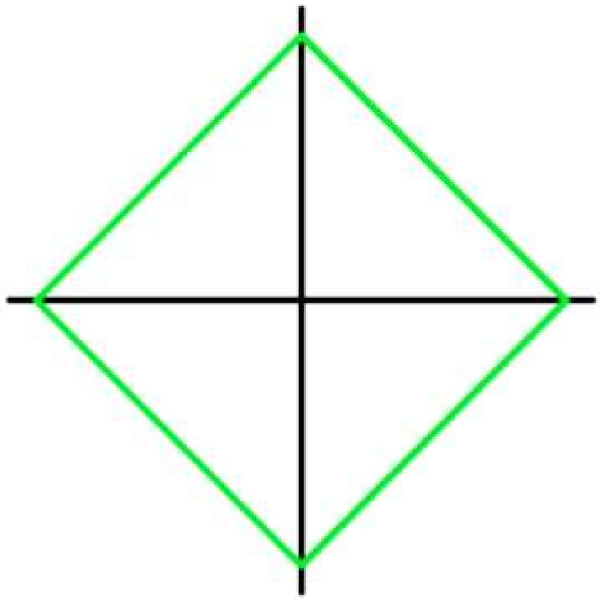
$$\mathbf{w}_{\mathrm{ML}} = \left(\lambda \mathbf{I} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^{\mathrm{T}}\mathbf{t}$$

We can also use a more general regularization function

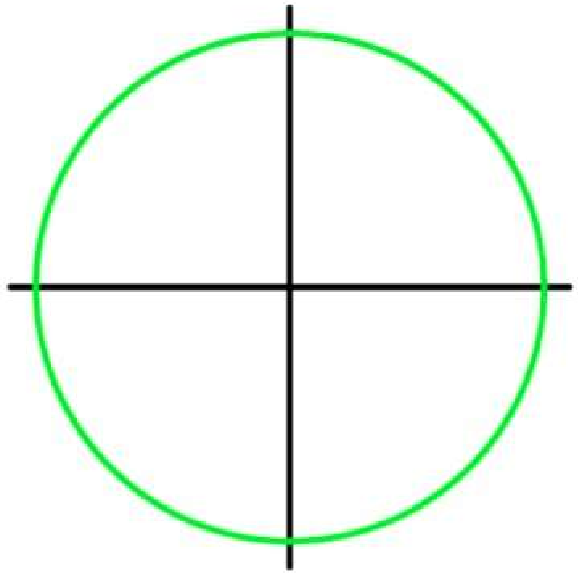
$$E_W(\mathbf{w}) = \frac{1}{2}\sum_{j=1}^{M} |w_j|^q$$

q=2 is the quadratic case shown already

Regularization Behavior

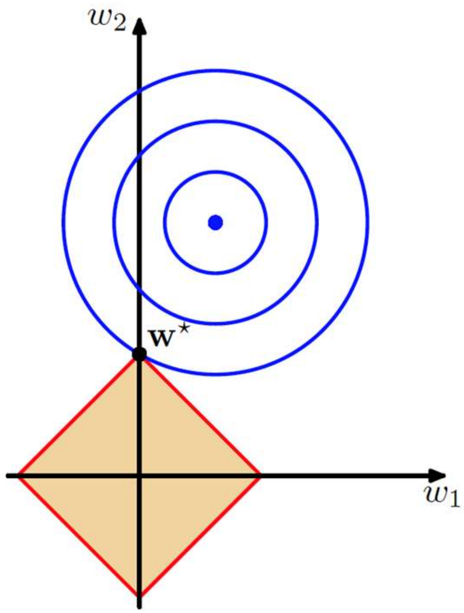


$q{=}1$ / L_1 / Lasso

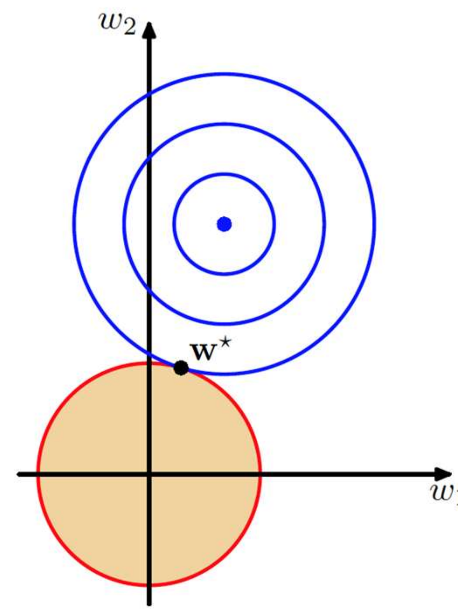


$q{=}2$ / L_2 / Ridge

Effect of Regularization



q=1 / L_1 / Lasso



q=2 / L_2 / Ridge

Bias vs. Variance

Expected squared loss:

$$E[L] = \int \{f(x) - h(x)\}^2 p(x)dx + \int \{h(x) - y\}^2 p(x,y)dxdy$$

- *f(x)* = prediction function
- *h(x)* = true function
- *y* = provided noisy value
- First term minimizes loss relative to the true model
- Second term minimizes error from noise

Bias vs. Variance

Imagine we can sample many datasets from the underlying distribution

Integrate the first term which represents accuracy of the model

- Data sample dependent due to the fitting of the prediction function based on $\mathcal{D}$

$$\int \{f(x|\mathcal{D}) - h(x)\}^2 \, p(x) dx$$

- What is the expectation of this term over many samples of $\mathcal{D}$?

Bias vs. Variance

$$E_D\left[(f(x|\mathcal{D}) - h(x))^2\right] = \text{algebra...}$$

bias^2

variance

$$= \{E_D[f(x|\mathcal{D})] - h(x)\}^2 +$$
$$E_D\left[\{f(x|\mathcal{D}) - E_D[f(x|\mathcal{D})]\}^2\right]$$

For learning we want to minimize this function

The result is a tradeoff between bias and variance

Parameter Tradeoff

Regularization parameter λ controls this tradeoff

Higher λ = more regularization

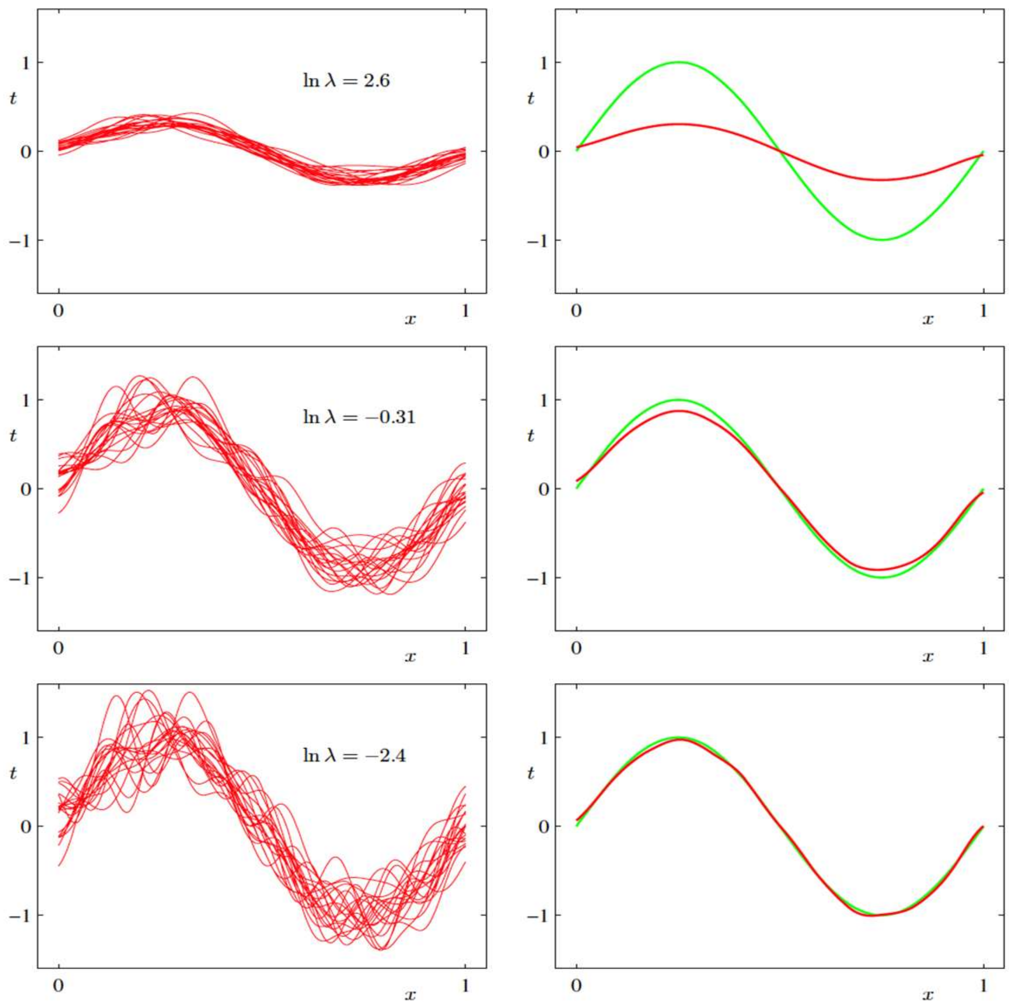
- Favors bias (under-fitting)

Lower λ = less regularization

- Favors variance (over-fitting)

More Regularization

Less Regularization



More Bias

More Variance

Problems

Maximum likelihood under-estimates variance and over-fits
- Try to fix using regularization

How do we decide model complexity?
- Parameter tuning on held out data

Is there a better way?
- Bayesian methods

Summary of Machine Learning Fundamentals

Fitting a function to data

Fitting: Optimization, what parameters can we change?

Function: Model, loss function

Data: Data/model assumptions? How we use data?

ML Algorithms: minimize a function on some data

Summary of Machine Learning Fundamentals

Data representation $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ $\mathbf{x}_i \in \Re^M$

Loss functions $y_i \in \Re$

Hypothesis class and tradeoffs

Generalization and bias/variance tradeoff

Learning settings: Supervised and unsupervised

Regularization

Sources of error

Next time: Classification!