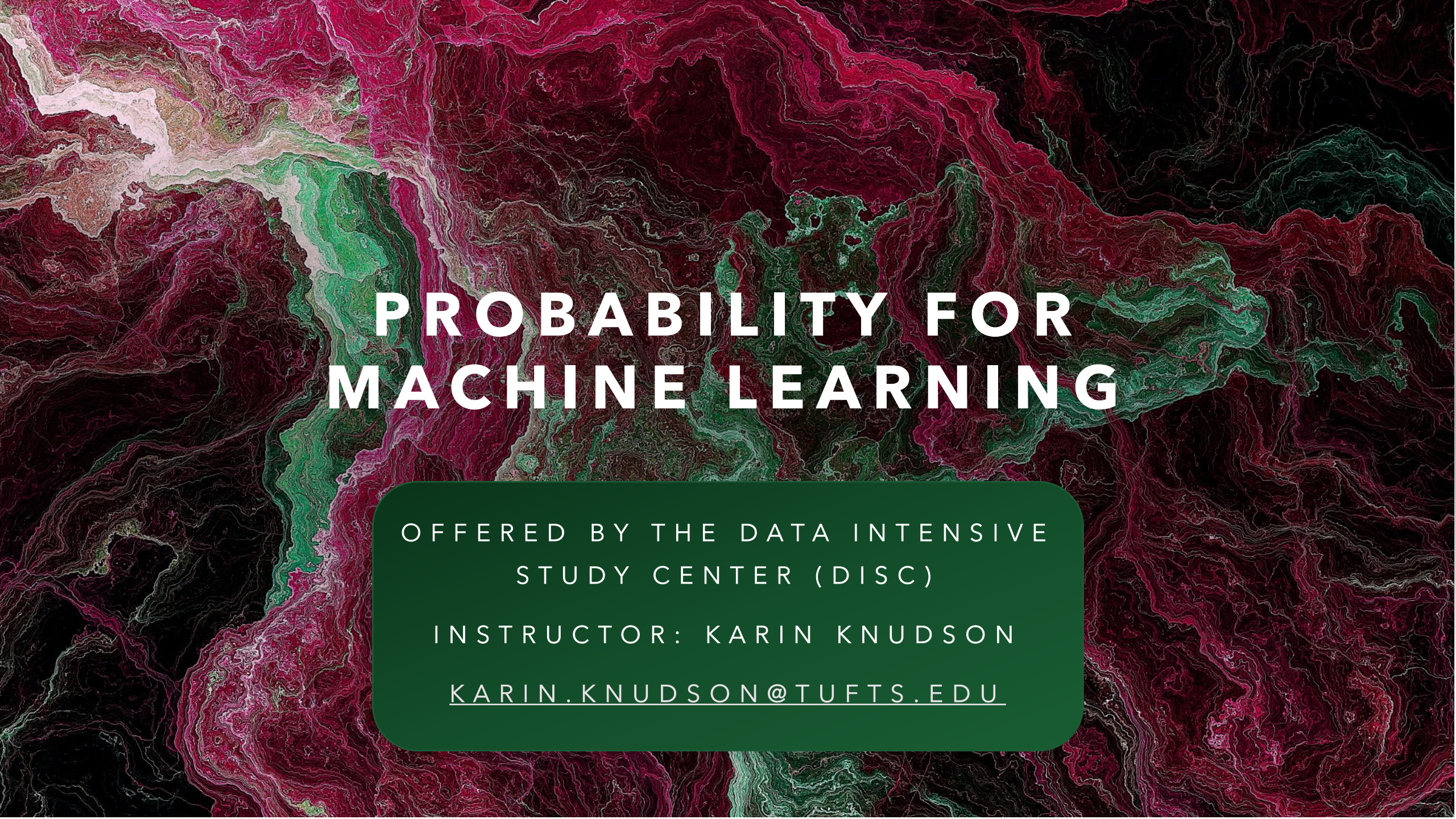


PROBABILITY FOR MACHINE LEARNING

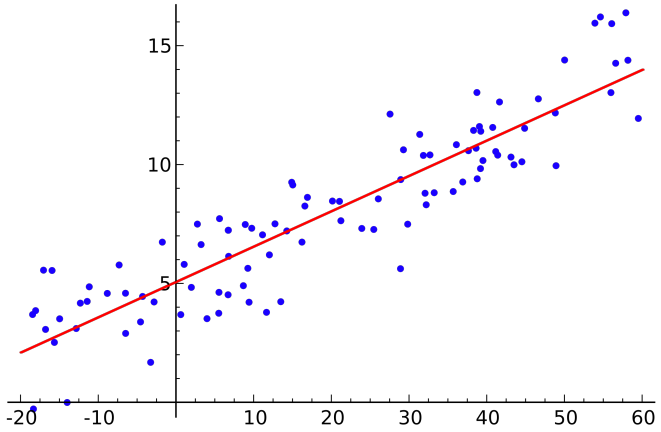
OFFERED BY THE DATA INTENSIVE STUDY CENTER (DISC)

INSTRUCTOR: KARIN KNUDSON

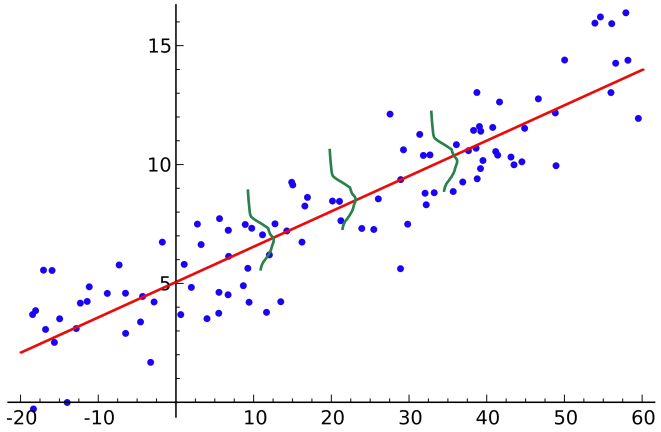
KARIN.KNUDSON@TUFTS.EDU

LINEAR REGRESSION

$$y = w_0 + w_1 X_1$$



LINEAR REGRESSION

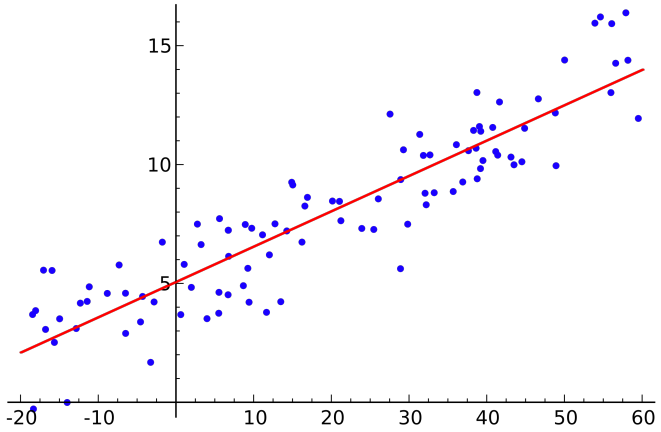


$$y = w_0 + w_1 X_1$$

$$y_i | X_i, w_0, w_1 \sim Normal(w_0 + w_1 X_1, \sigma^2)$$

$$p(y_i) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2} \right)$$

LINEAR REGRESSION



$$y = w_0 + w_1 X_1$$

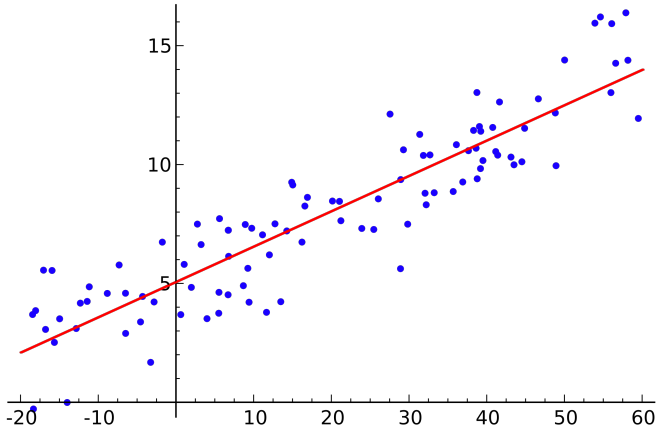
$$y_i | X_i, w_0, w_1 \sim Normal(w_0 + w_1 X_1, \sigma^2)$$

$$p(y_i) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right)$$

Likelihood

$$p(\mathbf{y}) = \prod_{i=1}^{N} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right)$$

LINEAR REGRESSION



$$y = w_0 + w_1 X_1$$

$$y_i | X_i, w_0, w_1 \sim Normal(w_0 + w_1 X_1, \sigma^2)$$

$$p(y_i) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right)$$

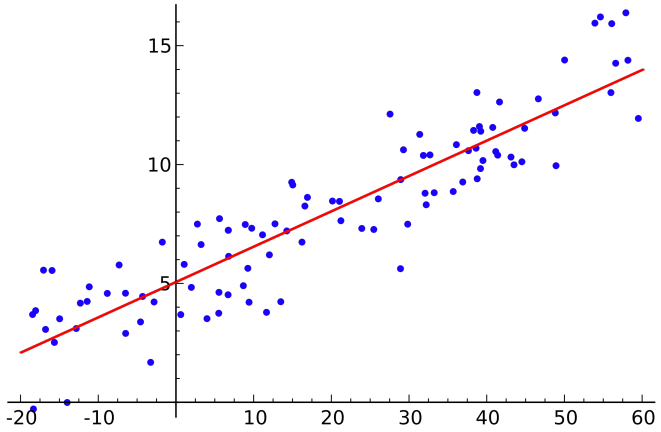
Likelihood

$$p(\mathbf{y}) = \prod_{i=1}^{N} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right)$$

$$\log p(\mathbf{y}|X, w) = -\sum_{i=1}^{N} \left(\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right) + \sum_{i=1}^{N} \log(1/(2\pi\sigma^2))$$

Log-likelihood (recall log(ab) = log(a) + log(b))

LINEAR REGRESSION



$$y = w_0 + w_1 X_1$$

$$y_i | X_i, w_0, w_1 \sim Normal(w_0 + w_1 X_1, \sigma^2)$$

$$p(y_i) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right)$$

Likelihood

$$p(\mathbf{y}) = \prod_{i=1}^{N} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right)$$

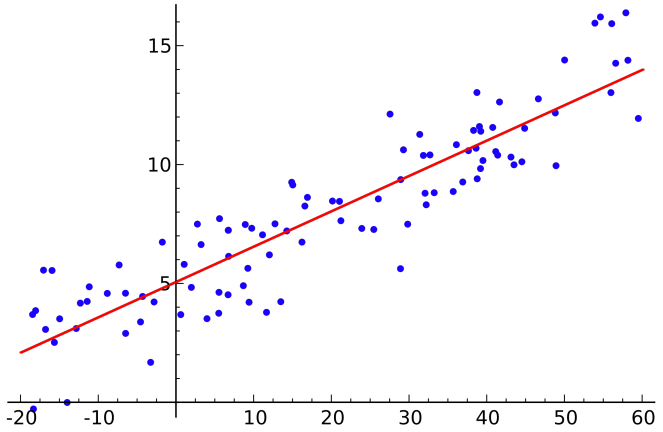
To maximize this with respect to w, it suffices to minimize:

$$\sum_{i=1}^{N} (y_i - (w_0 + w_1 X_i))^2$$

$$\log p(\mathbf{y} | X, w) = -\sum_{i=1}^{N} \left(\frac{(y_i - (w_0 + w_1 X_i))^2}{2\sigma^2}\right) + \sum_{i=1}^{N} \log(1/(2\pi\sigma^2))$$

Log-likelihood (recall log(ab) = log(a) + log(b))

LINEAR REGRESSION



- Log-likelihood easier to manage than the likelihood
- Maximize the likelihood by minimizing the **negative log-likelihood**
- Negative log-likelihood as the **loss function** that we seek to minimize
- Derived ordinary least squares setup from probabilistic assumptions about the data

LINEAR REGRESSION: INCLUDING BASIS FUNCTIONS

$$y = w_0 + w_1 x_1 + \ldots + w_D x_D$$

LINEAR REGRESSION: INCLUDING BASIS FUNCTIONS

$$y = w_0 + w_1 x_1 + \ldots + w_D x_D$$

Basis functions: give lots more flexibility!

$$y = w_0 + w_1 \phi_1(\mathbf{x}) + \ldots + w_{M-1} \phi_{M-1}(\mathbf{x})$$

$$y = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x})$$

Rewriting with vectors

$$y = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$$

$$\phi_0(\mathbf{x}) = 1$$

LINEAR REGRESSION: MAXIMUM LIKELIHOOD (AGAIN)

$$y_i|\mathbf{x}, \mathbf{w}, \sigma^2 \sim N(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i), \sigma^2)$$

LINEAR REGRESSION: MAXIMUM LIKELIHOOD (AGAIN)

$$y_i|\mathbf{x}, \mathbf{w}, \sigma^2 \sim N(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i), \sigma^2)$$

likelihood

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-(y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i))^2/2\sigma^2\right)$$

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \sigma^2) = \sum_{i=1}^{N} \left(-(y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i))^2/2\sigma^2\right) + \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}}$$

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i))^2 + N \log \frac{1}{\sqrt{2\pi\sigma^2}}$$

LINEAR REGRESSION: MAXIMUM LIKELIHOOD (AGAIN)

$$y_i|\mathbf{x}, \mathbf{w}, \sigma^2 \sim N(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i), \sigma^2)$$

likelihood

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-(y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i))^2/2\sigma^2\right)$$

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \sigma^2) = \sum_{i=1}^{N} \left(-(y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i))^2/2\sigma^2\right) + \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}}$$

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i))^2 + N \log \frac{1}{\sqrt{2\pi\sigma^2}}$$

Maximize likelihood by minimizing: $\sum_{i=1}^{N} (y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i))^2$

LINEAR REGRESSION: MAXIMUM LIKELIHOOD (AGAIN)

$$y_i|\mathbf{x}, \mathbf{w}, \sigma^2 \sim N(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i), \sigma^2)$$

likelihood

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-(y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i))^2/2\sigma^2\right)$$

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \sigma^2) = \sum_{i=1}^{N} \left(-(y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i))^2/2\sigma^2\right) + \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}}$$

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i))^2 + N \log \frac{1}{\sqrt{2\pi\sigma^2}}$$

Maximize likelihood by minimizing: $\sum_{i=1}^{N} (y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i))^2$

Moore-Penrose pseudo-inverse

$$\mathbf{w}_{ML} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y} = \Phi^\dagger\mathbf{y}$$

where $\Phi_{ij} = \phi_j(\mathbf{x}_i)$

LINEAR REGRESSION AND GRADIENT DESCENT

LINEAR REGRESSION AND REGULARIZATION: RIDGE REGRESSION

Same likelihood: $y_i|\mathbf{x}, \mathbf{w}, \sigma^2 \sim N(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i), \sigma^2)$

Can we put a prior over **w** (params of interest)?

LINEAR REGRESSION AND REGULARIZATION: RIDGE REGRESSION

Same likelihood: $y_i|\mathbf{x}, \mathbf{w}, \sigma^2 \sim N(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i), \sigma^2)$

Can we put a prior over **w** (params of interest)?

hyperparameter!

$w|\alpha \sim N(0, \alpha^2)$

LINEAR REGRESSION AND REGULARIZATION: RIDGE REGRESSION

likelihood $y_i|\mathbf{x}, \mathbf{w}, \sigma^2 \sim N(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i), \sigma^2)$

prior $w|\alpha \sim N(0, \alpha^2)$

LINEAR REGRESSION AND REGULARIZATION: RIDGE REGRESSION

likelihood $y_i|\mathbf{x}, \mathbf{w}, \sigma^2 \sim N(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i), \sigma^2)$

prior $w|\alpha \sim N(0, \alpha^2)$

posterior: $p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \sigma, \alpha) \propto \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-(y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i))^2/2\sigma^2\right) \prod_{j=0}^{M-1} \frac{1}{\sqrt{2\pi\alpha^2}} \exp(-w_j^2/2\alpha^2)$

LINEAR REGRESSION AND REGULARIZATION: RIDGE REGRESSION

likelihood $y_i|\mathbf{x}, \mathbf{w}, \sigma^2 \sim N(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i), \sigma^2)$

prior $w|\alpha \sim N(0, \alpha^2)$

posterior: $$p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \sigma, \alpha) \propto \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-(y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i))^2/2\sigma^2\right) \prod_{j=0}^{M-1} \frac{1}{\sqrt{2\pi\alpha^2}} \exp(-w_j^2/2\alpha^2)$$

$$\log p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \sigma, \alpha) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)) - \frac{1}{\alpha^2}\sum_{j=0}^{M-1} w_j^2 + C$$

LINEAR REGRESSION AND REGULARIZATION: RIDGE REGRESSION

likelihood $y_i|\mathbf{x}, \mathbf{w}, \sigma^2 \sim N(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i), \sigma^2)$

prior $w|\alpha \sim N(0, \alpha^2)$

posterior: $p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \sigma, \alpha) \propto \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-(y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i))^2/2\sigma^2\right) \prod_{j=0}^{M-1} \frac{1}{\sqrt{2\pi\alpha^2}} \exp(-w_j^2/2\alpha^2)$

$$\log p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \sigma, \alpha) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)) - \frac{1}{\alpha^2}\sum_{j=0}^{M-1} w_j^2 + C$$

Find MAP (=maximum a posteriori) estimate for **w** by minimizing:

$$\sum_{n=1}^{N}(y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)) + \frac{\sigma^2}{\alpha^2}\mathbf{w}^T\mathbf{w}$$

Minimize:

$$\sum_{n=1}^{N}(y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)) + \lambda||\mathbf{w}||^2$$

LINEAR REGRESSION AND REGULARIZATION: RIDGE REGRESSION

Minimize:

$$\sum_{n=1}^{N}(y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)) + \lambda||\mathbf{w}||^2$$

Regularization involves adding additional information to prevent overfitting.

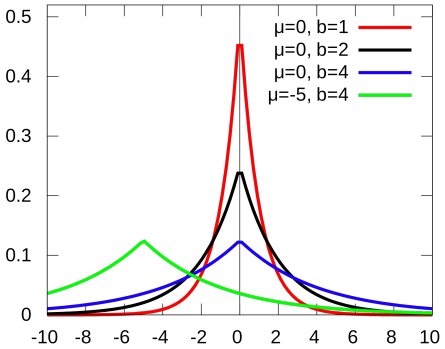
A common form it can take is adding a penalty term to the loss function.

LINEAR REGRESSION AND REGULARIZATION: LASSO

$$\sum_{n=1}^{N}(y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)) + \lambda||\mathbf{w}||_1^2$$

$$||\mathbf{w}||_1 := \sum |w_j|$$

Laplace distribution

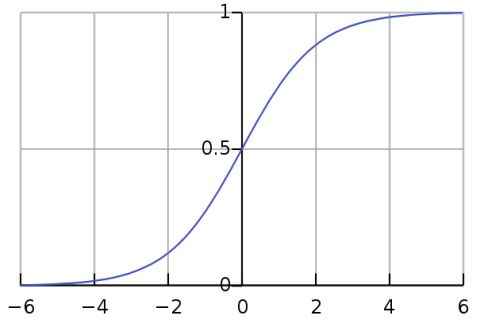


Can derive from putting a Laplacian prior over **w**

Leads to **sparse** solutions for **w** (i.e. many components of w are zero).

LOGISTIC REGRESSION FOR CLASSIFICATION

$$p(C_0|\phi) = y(\phi) = \sigma(\mathbf{w}^T\phi) = \frac{1}{1 + e^{-\mathbf{w}^T\phi}}$$



LOGISTIC REGRESSION FOR CLASSIFICATION

$$p(C_0|\phi) = y(\phi) = \sigma(\mathbf{w}^T\phi) = \frac{1}{1+e^{-\mathbf{w}^T\phi}}$$

Data: t =0 or 1, class 0 or class 1

$$p(t_i) = y_i^{t_i}(1-y_i)^{1-t_i}$$

$$p(\mathbf{t}|\mathbf{w}) = \prod_{i=1}^{N} y_i^{t_i}(1-y_i)^{1-t_i}$$

LOGISTIC REGRESSION FOR CLASSIFICATION

$$p(C_0|\phi) = y(\phi) = \sigma(\mathbf{w}^T\phi) = \frac{1}{1+e^{-\mathbf{w}^T\phi}}$$

Data: t =0 or 1, class 0 or class 1

$$p(t_i) = y_i^{t_i}(1-y_i)^{1-t_i}$$

$$p(\mathbf{t}|\mathbf{w}) = \prod_{i=1}^{N} y_i^{t_i}(1-y_i)^{1-t_i}$$

error function, which has a nice gradient: $\nabla E(\mathbf{w}) = \sum_{i=1}^{N}(y_n - t_n)\phi_n$

$$-\log p(\mathbf{t}|\mathbf{w}) = -\sum_{i=1}^{N} t_i \log y_i + (1-t_i)\log(1-y_i)$$

LOGISTIC REGRESSION FOR CLASSIFICATION

$$p(C_0|\phi) = y(\phi) = \sigma(\mathbf{w}^T\phi) = \frac{1}{1+e^{-\mathbf{w}^T\phi}}$$

Data: t =0 or 1, class 0 or class 1

$$p(t_i) = y_i^{t_i}(1-y_i)^{1-t_i}$$

$$p(\mathbf{t}|\mathbf{w}) = \prod_{i=1}^{N} y_i^{t_i}(1-y_i)^{1-t_i}$$

error function, which has a nice gradient: $\nabla E(\mathbf{w}) = \sum_{i=1}^{N}(y_n - t_n)\phi_n$

$$-\log p(\mathbf{t}|\mathbf{w}) = -\sum_{i=1}^{N} t_i \log y_i + (1-t_i)\log(1-y_i)$$

cross entropy

LOGISTIC REGRESSION FOR CLASSIFICATION

$$p(C_0|\phi) = y(\phi) = \sigma(\mathbf{w}^T\phi) = \frac{1}{1+e^{-\mathbf{w}^T\phi}}$$

Data: t =0 or 1, class 0 or class 1

$$p(t_i) = y_i^{t_i}(1-y_i)^{1-t_i}$$

$$p(\mathbf{t}|\mathbf{w}) = \prod_{i=1}^{N} y_i^{t_i}(1-y_i)^{1-t_i}$$

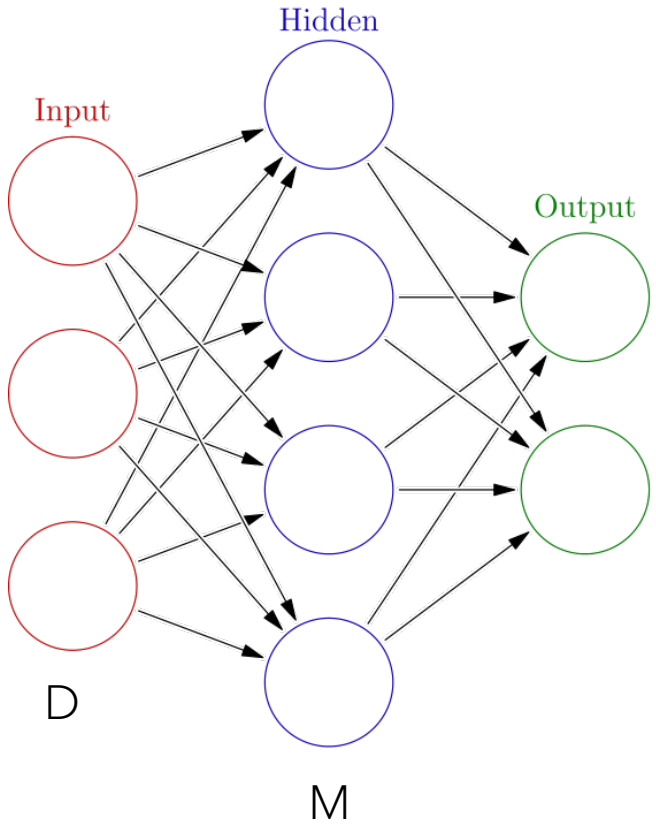
$$-\log p(\mathbf{t}|\mathbf{w}) = -\sum_{i=1}^{N} t_i \log y_i + (1-t_i)\log(1-y_i)$$

error function, which has a nice gradient: $\nabla E(\mathbf{w}) = \sum_{i=1}^{N}(y_n - t_n)\phi_n$

cross entropy

Note: we can regularize here too!
scikit-learn's logistic regression is regularized by default!

NEURAL NETWORKS



$$y_k(\mathbf{x}, \mathbf{w}) = \sigma\left(\sum_{j=1}^{M} w_{kj}^2 h\left(\sum_{i=1}^{D} w_{ji}^1 x_i + w_{j0}^1\right) + w_{k0}^2\right)$$

hidden unit

h some nonlinear function (e.g. logistic sigmoid, tanh)
x values are inputs

For binary classification: can consider error function as cross-entropy

$$-\log p(\mathbf{t}|\mathbf{w}) = -\sum_{i=1}^{N} t_i \log y_i + (1 - t_i) \log(1 - y_i)$$

DEEP LEARNING LOSS FUNCTIONS

A fascinating visualization project: https://losslandscape.com/