# 6.891 Machine learning and neural networks

Tommi Jaakkola
MIT AI LAB
*tommi@ai.mit.edu*

## Topics

- Linear regression, additive models

  Loss functions, fitting, generalization

  Statistical view
- Generalized linear/additive models

## Regression

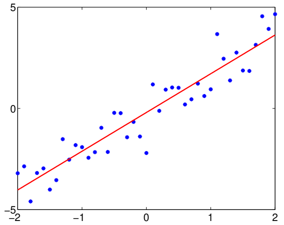
- We need to define a **function class** and **fitting criterion (loss)**
- Example: linear functions of one variable (two parameters)

$$f(x; \mathbf{w}) = w_0 + w_1 x \tag{1}$$

  with a squared loss: $\text{Loss}(y, f(x; \mathbf{w})) = (y - f(x; \mathbf{w}))^2/2$.

  Estimation based on minimizing the *empirical* loss

$$J_n(\mathbf{w}) = \sum_{i=1}^{n} \text{Loss}(y_i, f(x_i; \mathbf{w})) \tag{2}$$




## Linear regression: estimation

- We minimize the *empirical* squared loss

$$J_n(\mathbf{w}) = \sum_{i=1}^{n} \text{Loss}(y_i, f(x_i; \mathbf{w})) = \sum_{i=1}^{n} (y_i - w_0 - w_1 x_i)^2/2 \tag{3}$$

  Setting the derivatives with respect to $w_0$ and $w_1$ to zero we get necessary conditions for the "optimal" parameter values

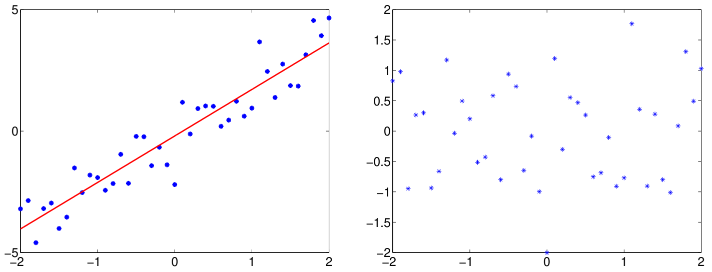
$$\frac{\partial}{\partial w_0} J_n(\mathbf{w}) = -\sum_{i=1}^{n} (y_i - w_0 - w_1 x_i) = 0 \tag{4}$$

$$\frac{\partial}{\partial w_1} J_n(\mathbf{w}) = -\sum_{i=1}^{n} (y_i - w_0 - w_1 x_i)\, x_i = 0 \tag{5}$$

  **Note:** These conditions mean that the prediction error $(y_i - w_0 - w_1 x_i)$ has zero mean and is decorrelated with the inputs $x_i$

## Linear regression: estimation

- The prediction error $(y_i - w_0 - w_1 x_i)$ is decorrelated with the inputs $x_i$



## Linear regression: estimation

$$\frac{\partial}{\partial w_0} J_n(\mathbf{w}) = -\sum_{i=1}^{n} (y_i - w_0 - w_1 x_i) = 0 \qquad (6)$$

$$\frac{\partial}{\partial w_1} J_n(\mathbf{w}) = -\sum_{i=1}^{n} (y_i - w_0 - w_1 x_i)\, x_i = 0 \qquad (7)$$

- Solution via matrix inversion

$$\begin{aligned} w_0 \left(\textstyle\sum_{i=1}^n 1\right) &+ w_1 \left(\textstyle\sum_{i=1}^n x_i\right) &= \textstyle\sum_{i=1}^n y_i \\ w_0 \left(\textstyle\sum_{i=1}^n x_i\right) &+ w_1 \left(\textstyle\sum_{i=1}^n x_i^2\right) &= \textstyle\sum_{i=1}^n y_i x_i \end{aligned}$$

or $\Phi \mathbf{w} = b$, where

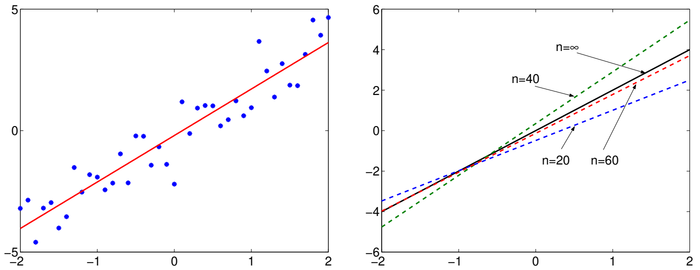
$$\Phi = \begin{bmatrix} \sum_{i=1}^n 1 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}, \quad b = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i x_i \end{bmatrix} \qquad (8)$$

- If $\Phi$ is invertible, we get our parameter estimates via $\hat{\mathbf{w}} = \Phi^{-1} b$

  What if $\Phi$ is not invertible?

## Linear regression: generalization

- Generalization performance as a function of the number of training examples $\{(x_1, y_1), \ldots, (x_n, y_n)\}$




- This makes no sense unless we assume that there is a systematic relation between $x$ and $y$: each training example $(x, y)$ is an *independent* sample from a fixed but unknown distribution $P(x, y)$.

## Linear regression: generalization

| | |
|---|---|
| Training examples | $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ |
| Test examples | $\{(x_{n+1}, y_{n+1}), \ldots, (x_{n+N}, y_{n+N})\}$ |

Let $\hat{\mathbf{w}}_n$ be the least squares parameter estimates on the basis of the training examples.

$$\text{Mean training error} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\mathbf{w}}_0 - \hat{\mathbf{w}}_1 x_i)^2$$

$$\text{Mean test error} = \frac{1}{N} \sum_{i=n+1}^{n+N} (y_i - \hat{\mathbf{w}}_0 - \hat{\mathbf{w}}_1 x_i)^2$$

$$\text{Generalization error} = E_{(x,y) \sim P} \left\{ (y - \hat{\mathbf{w}}_0 - \hat{\mathbf{w}}_1 x)^2 \right\}$$

- Can we ever compute the generalization error?

## Linear regression

- Linear in the parameters $\mathbf{w}$, not necessarily in the inputs $\mathbf{x}$
  1. Simple linear prediction $f : \mathcal{R} \to \mathcal{R}$
  $$f(x; \mathbf{w}) = w_0 + w_1 x \tag{9}$$
  2. $m^{th}$ order polynomial prediction $f : \mathcal{R} \to \mathcal{R}$
  $$f(x; \mathbf{w}) = w_0 + w_1 x + \ldots + w_{m-1} x^{m-1} + w_m x^m \tag{10}$$
  3. Multi-dimensional linear prediction $f : \mathcal{R}^d \to R$
  $$f(\mathbf{x}; \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_{d-1} x_{d-1} + w_d x_d \tag{11}$$
  where $\mathbf{x} = [x_1 \ldots x_{d-1}\; x_d]^T$, $d = dim(\mathbf{x})$

## Additive models

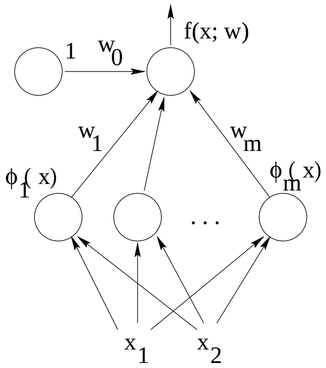
4. Prediction via linear combination of basis functions (features) $\{\phi_1(\mathbf{x}), \ldots, \phi_m(\mathbf{x})\}$, where each $\phi_i(\mathbf{x}) : \mathcal{R}^d \to \mathcal{R}$, and
$$f(\mathbf{x}; \mathbf{w}) = w_0 + w_1 \phi_1(\mathbf{x}) + \ldots + w_{m-1} \phi_{m-1}(\mathbf{x}) + w_m \phi_m(\mathbf{x}) \tag{12}$$

- For example:

  If $\phi_i(x) = x^i$, $i = 1, \ldots, m$, then
  $$f(x; \mathbf{w}) = w_0 + w_1 x + \ldots + w_{m-1} x^{m-1} + w_m x^m \tag{13}$$
  If $m = d$, $\phi_i(\mathbf{x}) = x_i$, $i = 1, \ldots, d$, then
  $$f(\mathbf{x}; \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_{d-1} x_{d-1} + w_d x_d \tag{14}$$

## Additive models

- Graphical representation of additive models




- What if the basis functions themselves can be adjusted?

## Additive models

- Example: we have $m$ prototypes of examples $\mu_1, \ldots, \mu_m$

  The basis functions can be used to (softly) select the closest prototype to each example $\mathbf{x}$
  $$\phi_k(\mathbf{x}) \;=\; \exp\{\, -\frac{1}{2}\|\mathbf{x} - \mu_k\|^2 \,\} \tag{15}$$
- Example: the "basis functions" may also be constructed by computing various relevant features from the examples
  $$\phi_k(\mathbf{x}) \;=\; \begin{cases} 1, \text{ if interest rate is up} \\ 0, \text{ otherwise} \end{cases} \tag{16}$$

## Estimating additive models

$$f(\mathbf{x};\mathbf{w}) = w_0 + w_1\phi_1(\mathbf{x}) + \ldots + w_{m-1}\phi_{m-1}(\mathbf{x}) + w_m\phi_m(\mathbf{x})$$

- Again, we can find the parameters by minimizing the *empirical loss*

$$J_n(\mathbf{w}) = \sum_{i=1}^{n} \text{Loss}(y_i, f(\mathbf{x}_i;\mathbf{w})) = \frac{1}{2}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i;\mathbf{w}))^2 \quad (17)$$

- Assuming a squared loss, the necessary conditions for optimal parameters are given by

$$\frac{\partial}{\partial w_0}J_n(\mathbf{w}) = -\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i;\mathbf{w})) = 0 \quad (18)$$

$$\ldots$$

$$\frac{\partial}{\partial w_m}J_n(\mathbf{w}) = -\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i;\mathbf{w}))\,\phi_m(\mathbf{x}_i) = 0 \quad (19)$$

  where as before the error $(y_i - f(\mathbf{x}_i;\mathbf{w}))$ has zero mean and is decorrelated with the basis functions $\{\phi_i(\mathbf{x}_i)\}$

## Estimation of additive models

- As before, we can rewrite the estimation equations

$$\begin{array}{lllll} w_0\left(\sum_{i=1}^{n} 1\right) + & \ldots & +w_m\left(\sum_{i=1}^{n}\phi_m(\mathbf{x}_i)\right) & = & \sum_{i=1}^{n} y_i \\ w_0\left(\sum_{i=1}^{n}\phi_1(\mathbf{x}_i)\right) + & \ldots & +w_m\left(\sum_{i=1}^{n}\phi_m(\mathbf{x}_i)\phi_1(\mathbf{x}_i)\right) & = & \sum_{i=1}^{n} y_i\phi_1(\mathbf{x}_i) \\ & & & \ldots & \\ w_0\left(\sum_{i=1}^{n}\phi_m(\mathbf{x}_i)\right) + & \ldots & +w_m\left(\sum_{i=1}^{n}\phi_m(\mathbf{x}_i)\phi_m(\mathbf{x}_i)\right) & = & \sum_{i=1}^{n} y_i\phi_m(\mathbf{x}_i) \end{array}$$

  in a matrix form $\Phi\mathbf{w} = b$ and solve via (pseudo-)inverse

  (what about *regularization*?)

## Statistical view of linear regression

- A statistical regression model

  **Observed output = function + noise**

$$y = f(\mathbf{x};\mathbf{w}) + \epsilon \quad (20)$$

  where, e.g., $\epsilon \sim N(0,\sigma^2)$.

- Whatever we cannot capture with our chosen family of functions will be (assumed) noise
- Our function $f(\mathbf{x};\mathbf{w})$ is trying to capture the mean of the observations $y$ given $\mathbf{x}$:

$$E\{\,y\,|\,\mathbf{x}\,\} = f(\mathbf{x};\mathbf{w}) \quad (21)$$

## Statistical view of linear regression

- According to our statistical model

$$y = f(\mathbf{x};\mathbf{w}) + \epsilon,\ \ \epsilon \sim N(0,\sigma^2) \quad (22)$$

  the outputs $y$ given $\mathbf{x}$ are normally distributed with mean $f(\mathbf{x};\mathbf{w})$ and variance $\sigma^2$:

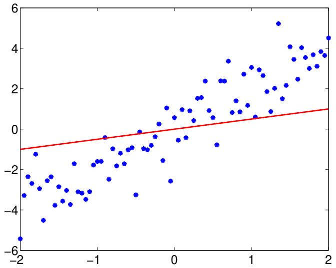
$$P(y|\mathbf{x},\mathbf{w},\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2\sigma^2}(y - f(\mathbf{x};\mathbf{w}))^2\} \quad (23)$$

- As a result we can evaluate the uncertainty in the predictions, not just the mean
- Loss function? Estimation?

## Maximum likelihood estimation

- Given observations $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ we find the parameters $\mathbf{w}$ that maximize the likelihood of the observed outputs

$$L(D; \mathbf{w}, \sigma^2) = \prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w}, \sigma^2) \qquad (24)$$




Why is this a bad fit according to the likelihood criterion?

## Maximum likelihood estimation

Likelihood of the observed outputs:

$$L(D; \mathbf{w}, \sigma^2) = \prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w}, \sigma^2) \qquad (25)$$

- It is often easier (but equivalent) to try to maximize the log-likelihood:

$$\begin{aligned} l(D; \mathbf{w}, \sigma^2) &= \log L(D; \mathbf{w}, \sigma^2) = \sum_{i=1}^{n} \log P(y_i|\mathbf{x}_i, \mathbf{w}, \sigma^2) && (26) \\ &= \sum_{i=1}^{n} \left( -\frac{1}{2\sigma^2}(y_i - f(\mathbf{x}_i; \mathbf{w}))^2 - \log \sqrt{2\pi\sigma^2} \right) && (27) \\ &= \left( -\frac{1}{2\sigma^2} \right) \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i; \mathbf{w}))^2 + \ldots && (28) \end{aligned}$$

- This should look familiar...

## Maximum likelihood estimation cont'd

- The noise distribution and the loss-function are intricately related

$$\text{Loss}(y, f(\mathbf{x}; \mathbf{w})) = -\log P(y|\mathbf{x}, \mathbf{w}, \sigma^2) + \text{const.} \qquad (29)$$

## Maximum likelihood estimation cont'd

- General fitting criterion: likelihood of the observed outputs

$$L(D; \mathbf{w}, \sigma^2) = \prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w}, \sigma^2) \qquad (30)$$

- We can just as easily fit the noise variance $\sigma^2$ by maximizing the log-likelihood $l(D; \mathbf{w}, \sigma^2)$ with respect to $\sigma^2$

  What might the answer be?

## Maximum likelihood estimation cont'd

- General fitting criterion: likelihood of the observed outputs

$$L(D; \mathbf{w}, \sigma^2) = \prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w}, \sigma^2) \tag{31}$$

- We can just as easily fit the noise variance $\sigma^2$ by maximizing the log-likelihood $l(D; \mathbf{w}, \sigma^2)$ with respect to $\sigma^2$

  Let $\hat{\mathbf{w}}$ be the maximum likelihood parameters for the linear model $f(\mathbf{x}; \mathbf{w})$, we can compute $\sigma^2$ as

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i; \hat{\mathbf{w}}))^2 \tag{32}$$

  i.e., it is the mean squared prediction error of the best linear predictor.

## From regression to classification

Linear predictor, e.g., $f(x; \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2$

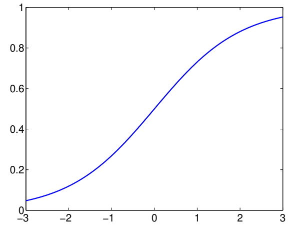
- *Linear regression model*:

$$E\{\, y \,|\, x \,\} = f(x; \mathbf{w}) \tag{33}$$

- *Generalized linear model*:

$$E\{\, y \,|\, x \,\} = g\,(\, f(x; \mathbf{w})\,) \tag{34}$$

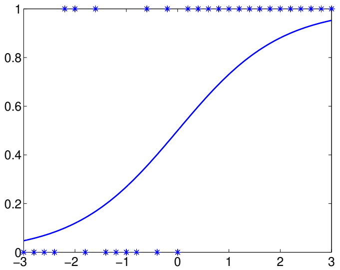
where $g(\cdot)$ is an invertible (squashing) function such as the logistic function $g(z) = (1 + \exp(-z))^{-1}$




## Logistic regression

- Instead of trying to predict the class label $y \in \{0, 1\}$ directly, we model the probability of the label given the input example $x$

  For binary variable $y$, $E\{\, y \,|\, x \,\} = P(y = 1|x)$




$$P(y = 1|x, \mathbf{w}) = g\,(\, w_0 + w_1 x_1\,) \tag{35}$$