Introduction to Probabilistic Machine Learning

Ralf Herbrich, Rainer Schlosser

Inference & Decision Making

Overview

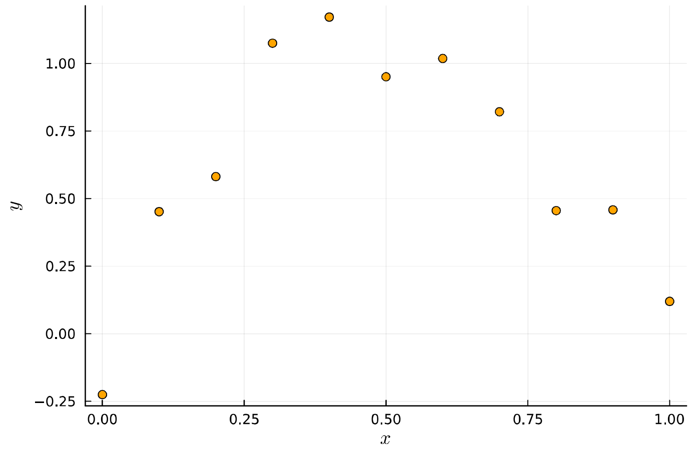
1. Inference Methods
   - Bayesian Inference
   - Maximum Likelihood Estimation
2. Decision Making

Overview

1. **Inference Methods**
   - Bayesian Inference
   - Maximum Likelihood Estimation
2. Decision Making

Probabilistic Machine Learning: Ingredients

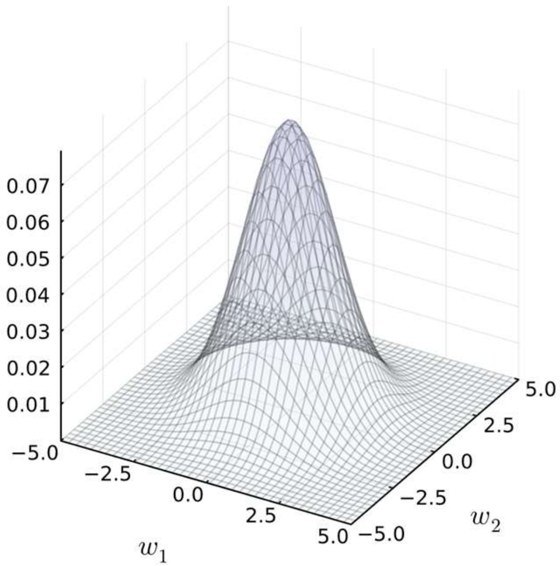
1. **Training Data**: $D \in (\mathcal{X} \times \mathcal{Y})^n$ of n (labelled) examples from the input space $\mathcal{X}$ and output space $\mathcal{Y}$



Training Data

$D \subset \mathbb{R}^2$

2. **Prior belief over functions from $\mathcal{X}$ to $\mathcal{Y}$**: $p(f), \quad f \in \mathcal{F}$
   - Space of functions, $\mathcal{F}$, is also called *hypothesis space*.

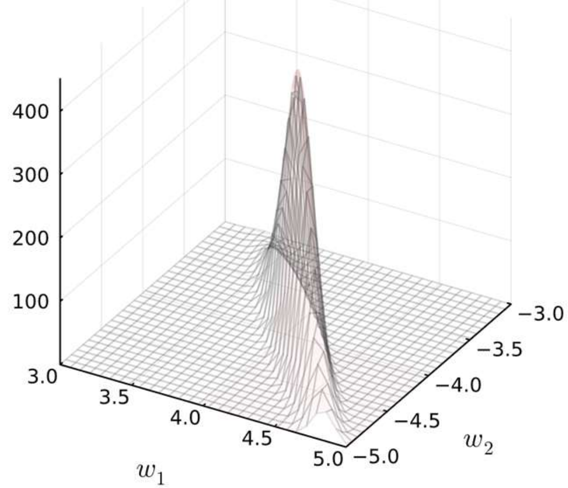


$$f_{\mathbf{w}}(x) = w_1 \cdot x + w_2 \cdot x^2$$

Prior

$$p(f_{\mathbf{w}}) = \mathcal{N}(w_1; 0,1) \cdot \mathcal{N}(w_2; 0,1)$$

3. **Likelihood of function**: $p(D|f) =: \ell(f)$
   - Link between data and functions
   - Models all assumptions how data/labels are generated from a function



Likelihood

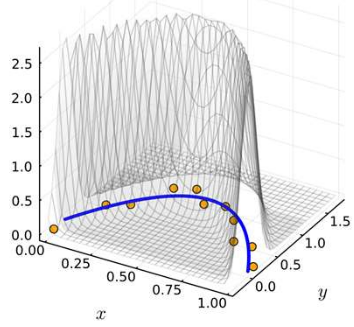
$$\ell\left(f_{\mathbf{w}}\right) = \prod_i \mathcal{N}\left(y_i; w_1 x_i + w_2 x_i^2, \sigma^2\right)$$

Key Question I: Inference and Predictive Distribution

Total probability rule

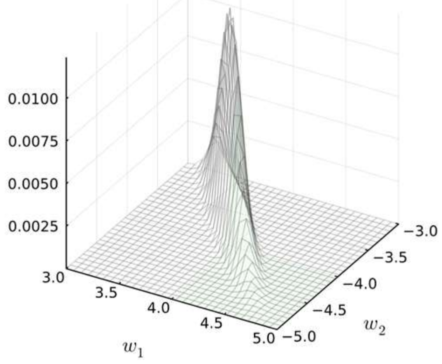
- **Predictive Distribution**. *Given a training set $D \in (\mathcal{X} \times \mathcal{Y})^n$ and a new input point $x \in \mathcal{X}$, the distribution $p(y|x, D) = \int p(y|x, f) \cdot p(f|D)\, df$ of target values $y \in \mathcal{Y}$ at the input point x is called the predictive distribution.*
  - **Observation 1**: For any function $f \in \mathcal{F}$ from the hypothesis space, the likelihood is already the distribution $p(y|x, f)$!
  - **Observation 2**: Each function $f \in \mathcal{F}$ from the hypothesis space has a posterior belief $p(f|D)$ after we have observed the training data using Bayes' rule!

$P(y|x, f_{\mathbf{w}})$

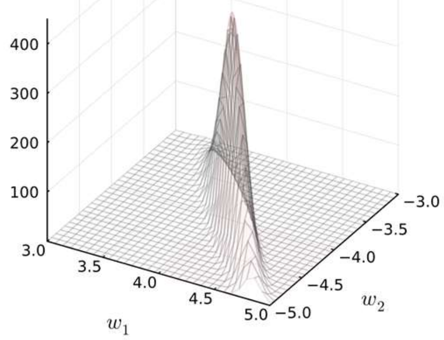


$$p(f|D) = \frac{p(D|f) \cdot p(f)}{p(D)}$$

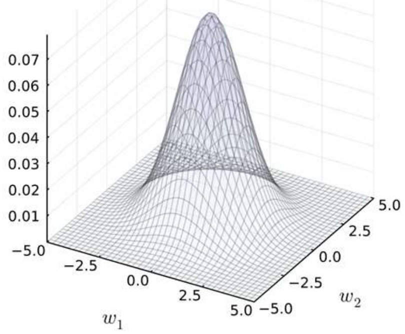
$P(f_{\mathbf{w}}|D)$



$P(D|f_{\mathbf{w}})$



$P(f_{\mathbf{w}})$



Overview

1. Inference Methods
   - **Bayesian Inference**
   - Maximum Likelihood Estimation
2. Decision Making

- **Two computational difficulties**:
  1. **Posterior** $p(f|D)$ requires the *multiplication* of likelihood with prior which often results in a distribution which is no longer in a family with very few parameters.

$$p(f|D) = \frac{p(D|f) \cdot p(f)}{p(D)} \propto \ell(f) \cdot p(f)$$

  2. **Predictive distribution** $p(y|x, D)$ requires the *summation* of the data distribution over all prediction functions. This is only feasible for a small number of parametric distributions.

$$p(y|x, D) = \int p(y|x, f) \cdot p(f|D)\, df$$

Probability Distributions: Conjugacy

- **Bayes Rule for Random Variables**. *For any probability distribution p over two random variables X and Θ, it holds*

$$\underbrace{p(\theta|x)}_{\text{Posterior}} = \frac{\overbrace{\boxed{\overbrace{p(x|\theta)}^{\text{Likelihood}} \cdot \underbrace{p(\theta)}_{\text{Prior}}}}^{p(x,\theta)}}{p(x)}$$

- **Conjugacy**. *A family $\{p(x,\theta)\}_{x,\theta}$ is conjugate if the posterior $p(\theta|x)$ is part of the same family as the prior $p(\theta)$ for any value of x.*

Likelihood $p(x|\theta)$	**Prior** $p(\theta)$	**Posterior** $p(\theta|x)$
$\mathrm{Ber}(x;\theta)$	$\mathrm{Beta}(\theta;\alpha,\beta)$	$\mathrm{Beta}\big(\theta;\alpha+x,\beta+(1-x)\big)$
$\mathrm{Bin}(x;n,\theta)$	$\mathrm{Beta}(\theta;\alpha,\beta)$	$\mathrm{Beta}\big(\theta;\alpha+x,\beta+(n-x)\big)$
$\mathcal{N}(x;\theta,\sigma^2)$	$\mathcal{N}(\theta;m,s^2)$	$\mathcal{N}\left(\theta; x\cdot\frac{s^2}{s^2+\sigma^2}+m\cdot\frac{\sigma^2}{s^2+\sigma^2}, s^2\cdot\frac{\sigma^2}{s^2+\sigma^2}\right)$

- **Big Advantage**: Computing the exact posterior is computationally efficient!

Howard Raiffa (1924 – 2016)

Robert Osher Schlaifer (1914 – 1994)

Introduction to Probabilistic Machine Learning

Unit 2 - Inference & Decision Making

Normal Distribution: Representations

- **Scale-Location Parameters**

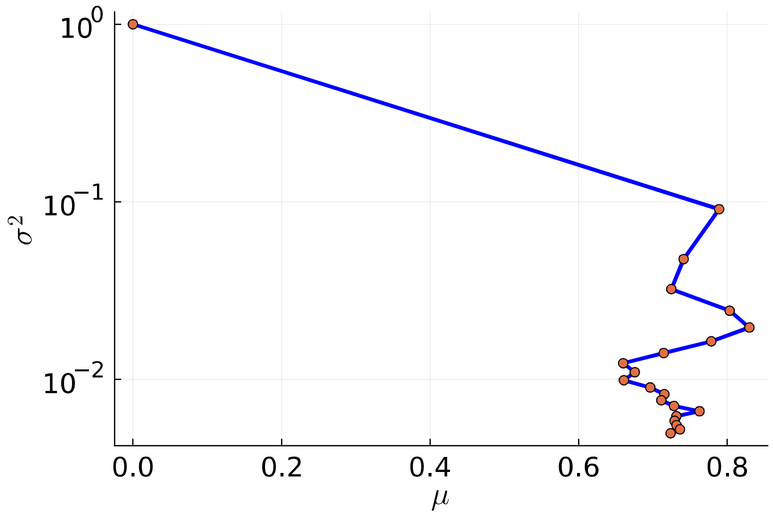
$$\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- **Conversions**

$$\mathcal{N}(x;\mu,\sigma^2) = \mathcal{G}\left(x;\frac{\mu}{\sigma^2},\frac{1}{\sigma^2}\right)$$

Two divisions only!

- **Posterior Inference**



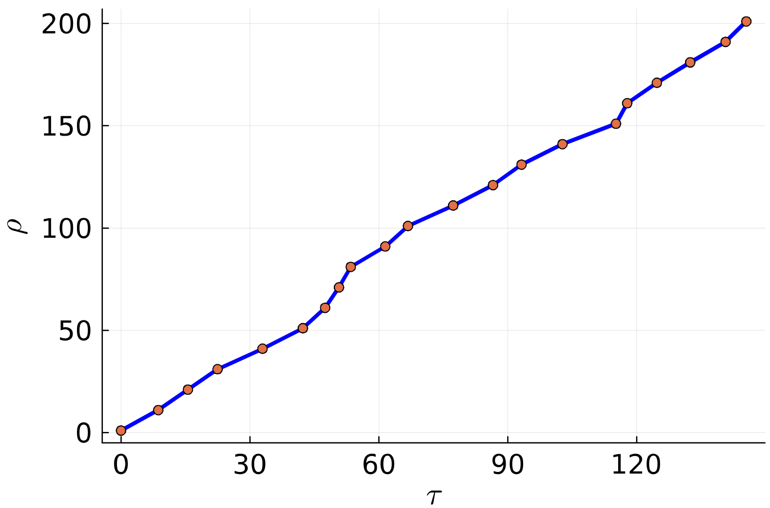
- **Natural Parameters**

$$\mathcal{G}(x;\tau,\rho) = \sqrt{\frac{\rho}{2\pi}} \cdot \exp\left(-\frac{\tau^2}{2\rho}\right) \cdot \exp\left(\tau \cdot x - \rho \cdot \frac{x^2}{2}\right)$$

- **Conversions**

$$\mathcal{G}(x;\tau,\rho) = \mathcal{N}\left(x;\frac{\tau}{\rho},\frac{1}{\rho}\right)$$

- **Posterior Inference**



Normal Distributions: Efficient Products & Divisions

- **Theorem (Multiplication)**. *Given two one-dimensional Gaussian distributions* $\mathcal{G}(x;\tau_1,\rho_1)$ *and* $\mathcal{G}(x;\tau_2,\rho_2\ )$ *we have*

Gaussian density

$$\mathcal{G}(x;\tau_1,\rho_1)\cdot\mathcal{G}(x;\tau_2,\rho_2)=\mathcal{G}(x;\tau_1+\tau_2,\rho_1+\rho_2)\cdot\mathcal{N}(\mu_1;\mu_2,\sigma_1^2+\sigma_2^2)$$

Additive updates!

- **Theorem (Division)**. *Given two one-dimensional Gaussian distributions* $\mathcal{G}(x;\tau_1,\rho_1)$ *and* $\mathcal{G}(x;\tau_2,\rho_2)$ *where* $\rho_1\geq\rho_2$ *we have*

Correction factor

$$\frac{\mathcal{G}(x;\tau_1,\rho_1)}{\mathcal{G}(x;\tau_2,\rho_2)}=\frac{\mathcal{G}(x;\tau_1-\tau_2,\rho_1-\rho_2)}{\mathcal{N}(\mu_1;\mu_2,\sigma_2^2-\sigma_1^2)}\cdot\frac{\sigma_2^2}{\sigma_2^2-\sigma_1^2}$$

Subtractive updates!

Gaussian density

Limit Normal Distributions: Dirac Delta and Uniform

- **Dirac Delta**. The Dirac delta function $\delta(\cdot)$ is defined as the limit $\sigma^2 \to 0$

$$\delta(x) = \lim_{\sigma^2 \to 0} \mathcal{N}(x; 0, \sigma^2)$$

- **Gaussian Uniform**. The Gaussian uniform $\mathcal{U}(\cdot)$ is defined as the limit $\sigma^2 \to \infty$

$$\mathcal{U}(x) = \lim_{\sigma^2 \to +\infty} \mathcal{N}(x; 0, \sigma^2)$$

- **Theorem (Convolution of Normal with Dirac)**. *For any $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$*

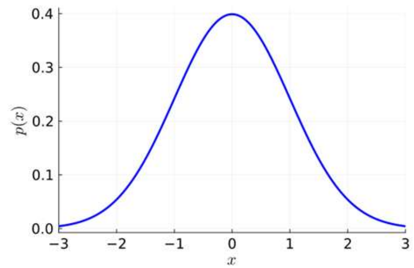
$$\int_{-\infty}^{+\infty} \delta(x) \cdot \mathcal{N}(x; \mu, \sigma^2)\, \mathrm{d}x = \boxed{\mathcal{N}(0; \mu, \sigma^2)}$$

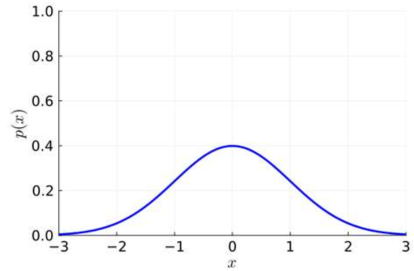
Gaussian density at $x = 0$

- **Theorem (Product of Normal with Uniform)**. *For any $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$*

$$\frac{\mathcal{U}(x) \cdot \mathcal{N}(x; \mu, \sigma^2)}{\int_{-\infty}^{+\infty} \mathcal{U}(\tilde{x}) \cdot \mathcal{N}(\tilde{x}; \mu, \sigma^2)\, \mathrm{d}\tilde{x}} = \boxed{\mathcal{N}(x; \mu, \sigma^2)}$$

Equivalent to multiplying with 1





Probability Distributions: Exponential Family

- **Exponential Family**. *A family of distributions is said to belong to the exponential family if the probability density/mass function in terms of the parameterisation* $\boldsymbol{\theta}$ *is*

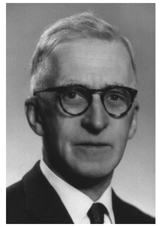
$$p(x) = \exp\left(\sum_i \eta_i(\boldsymbol{\theta}) \cdot T_i(x) - A(\boldsymbol{\theta})\right)$$

  - The η_i's are called canonical parameters and the T_i's are called sufficient statistics.

Distribution $p(x)$	Canonical Parameters $\eta(\theta)$	Sufficient Statistic $T(x)$
$\mathrm{Bin}(x; n, \pi)$	$\log\left(\frac{\pi}{1-\pi}\right)$	x
$\mathrm{Beta}(\pi; \alpha, \beta)$	$[\alpha, \beta]$	$[\log(\pi), \log(1-\pi)]$
$\mathcal{N}(x; \mu, \sigma^2)$	$\left[\frac{\mu}{\sigma^2}, \frac{1}{\sigma^2}\right]$	$\left[x, -\frac{x^2}{2}\right]$

- **Big Advantage**: Closed and efficient under multiplication (Bayes' rule!)

$$p(x; \boldsymbol{\eta}_1) \cdot p(x; \boldsymbol{\eta}_2) = p(x; \boldsymbol{\eta}_1 + \boldsymbol{\eta}_2)$$


Edwin Pitman (1897 - 1993)


Georges Darmois (1888 - 1960)


Bernhard Koopman (1900 - 1991)

Overview

1. Inference Methods
   - Bayesian Inference
   - **Maximum Likelihood Estimation**
2. Decision Making

Maximum Likelihood/Maximum A-Posteriori Inference

1. **Maximum Likelihood**. Find the most *likely* function $f_{\mathrm{ML}}(D)$ given the data D and approximate $p(f|D)$ by a single point distribution around
$$f_{\mathrm{ML}}(D) = \underset{f}{\operatorname{argmax}}\, p(D|f)$$
2. **Maximum A Posterior**. Find the most *probable* function $f_{\mathrm{MAP}}(D)$ given the data D and prior $p(f)$ and approximate $p(f|D)$ by a single point distribution around
$$f_{\mathrm{MAP}}(D) = \underset{f}{\operatorname{argmax}}\, p(D|f) \cdot p(f)$$

- **Pros**:
  1. Learning = optimization in the hypothesis space ("gradient descent")
  2. Storing the model = storing the function parameters
- **Cons**:
  1. The posterior/likelihood is "peaked" around a single best predictor (convergence)
  2. No model uncertainty after learning from data

Sir Ronald Fisher
(1890 – 1962)

Newton-Raphson Algorithm

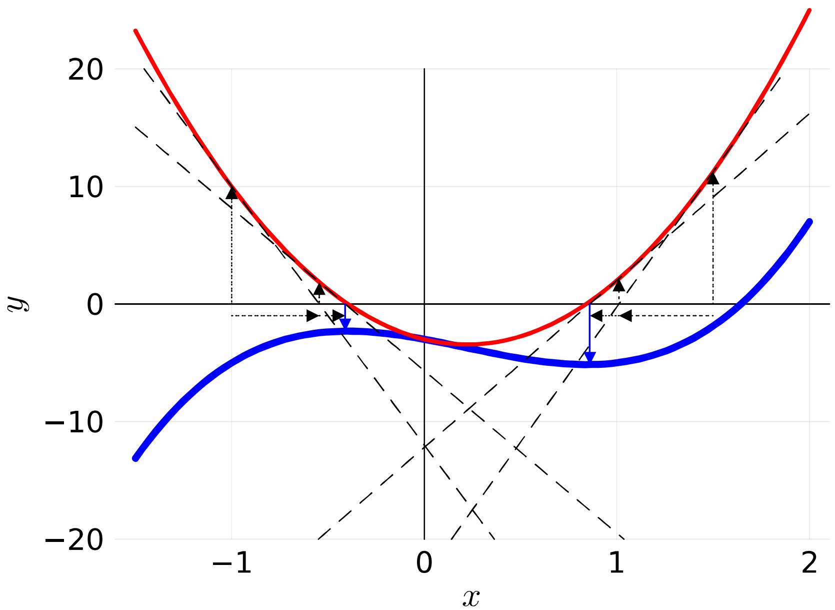
- **Problem**: Find the local extrema of a function $f: \mathbb{R} \to \mathbb{R}$
- **Idea**: Find the zeros of the first derivative f' of the function!
- **Newton-Raphson Algorithm**: Approximate f' at a point x_t with a linear function $g(x) = ax + b$ and find update x_{t+1} such that $g(x_{t+1}) = 0$

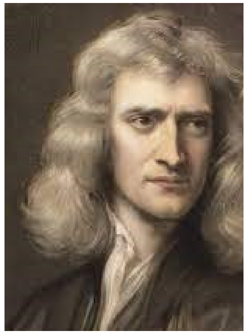
$$a = f''(x_t)$$

$$b = f'(x_t) - f''(x_t) \cdot x_t$$

$$x_{t+1} = -\frac{b}{a} = \frac{f''(x_t) \cdot x_t - f'(x_t)}{f''(x_t)}$$

$$\boxed{x_{t+1} = x_t - \frac{f'(x_t)}{f''(x_t)}}$$

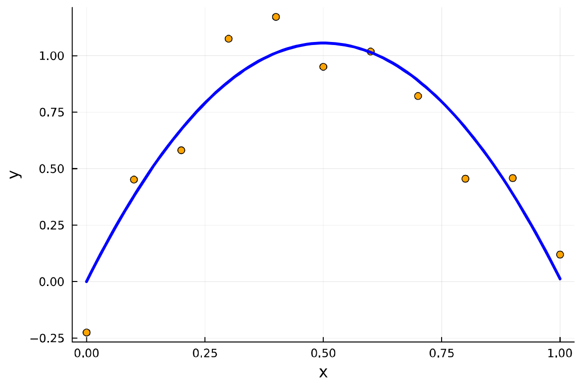




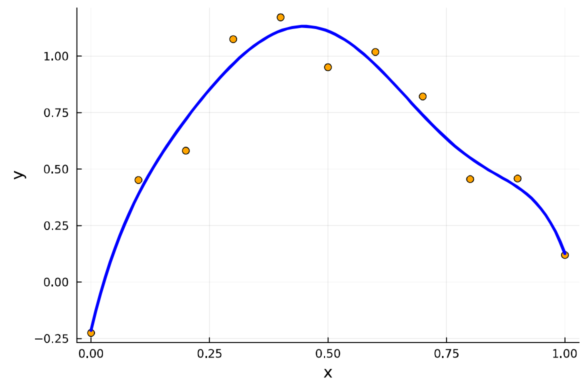
Sir Isaac Newton (1643 – 1727)

Maximum A-Posteriori Inference: Polynomial Regression

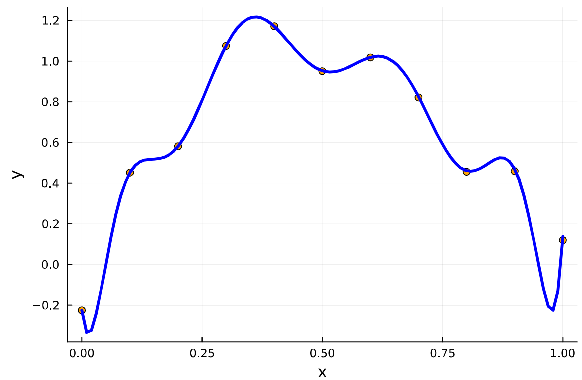
$$f(x) = w_1 x + w_2 x^2$$



$$f(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5 + w_6 x^6$$

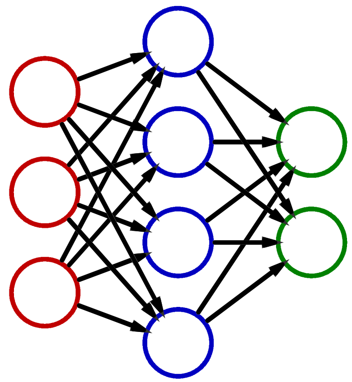


$$f(x) = \sum_{i=0}^{10} w_i \cdot x^i$$



Relation to Deep Learning

- **Deep Learning** is maximum likelihood inference on a layered function model
  - **Neural Networks**: $f(\boldsymbol{x}) = h(\boldsymbol{W}_L \cdots h(\boldsymbol{W}_2 h(\boldsymbol{W}_1 \boldsymbol{x})))$ where h is a sigmoid
    - Number of layers: L
    - Each element of each vector is called a “neuron”
    - Each product of the inner products is called a “synapse”



- **Maximum Likelihood** optimization via gradient descent (w.r.t. $\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_L$)
  - Application of the chain rule of differentiation = back propagation
  - Predicting and gradient computations are matrix multiplications; today, they are sped up using GPUs (which parallelize matrix multiplication)

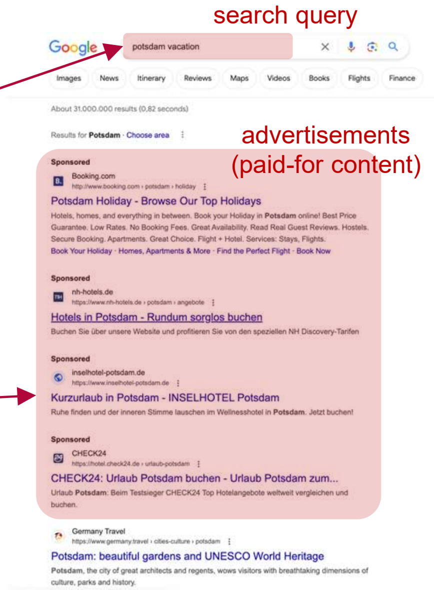
- **Regularization** for the Deep Learning algorithms are equivalent to prior assumptions on $p(\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_L)$!

Overview

1. Inference Methods
   - Bayesian Inference
   - Maximum Likelihood Estimation
2. **Decision Making**

Key Question II: Decision Making and Prediction

- **Decision Making**. *Given a training set $D \in (\mathcal{X} \times \mathcal{Y})^n$, a new input point $x \in \mathcal{X}$ and an action space $\mathcal{A}$, what action $\hat{a} \in \mathcal{A}$ shall be made at x based on D?*
  - **Example**: Deciding which of n advertisement to show on a search result page
    - x_i contains all information of all advertisements i, the search query, the user, …
    - $y_i \in \{0,1\}$ indicates, whether or not the advertisement is clicked by the user
    - D is the dataset of all displayed advertisements and whether or not they got clicked
    - $a \in \{1, \dots, n\}$ indicates, which one of the advertisements got chosen
  - **Decision theory** is concerned with the theory of making decisions based on uncertain outcomes and assigning numerical consequences to the outcome
- **Prediction**. *Given a training set $D \in (\mathcal{X} \times \mathcal{Y})^n$ and a new input point $x \in \mathcal{X}$, what prediction $\hat{y} \in \mathcal{Y}$ shall be made for the example x based on D?*
  - **Observation 1**: Special case of decision making when $\mathcal{A} = \mathcal{Y}$
  - **Observation 2**: In contrast to the *predictive distribution* $p(y|x, D)$ over **all** possible outcomes $y \in \mathcal{Y}$, we are committing to a **specific** outcome $\hat{y}(x, D) \in \mathcal{Y}$ in *prediction*!



Loss Functions

- **Problem**: What is the consequence of each action $\hat{a} \in \mathcal{A}$ or $\hat{y} \in \mathcal{Y}$ given that the truth was $y \in \mathcal{Y}$?
- **Loss Function**. *A loss function $l: \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$ is a function mapping the outcome space $\mathcal{Y}$ and an action space $\mathcal{A}$ to a real number representing the "cost" associated with taking the action $a \in \mathcal{A}$ when the true state of the world is $y \in \mathcal{Y}$.*
  - Losses are given by the domain problem; there are no "true" losses!
  - **Decision Making Example 1**: Deciding which of n advertisement to show
    - When advertisement $a \in \{1, \ldots, n\}$ is chosen and the click happens ($y_a = 1$), then the utility is the bid amount b_a being paid; all other advertisers do not pay

$$l(y, a) = -y_a \cdot b_a$$

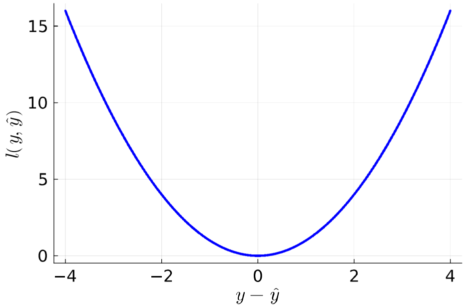
  - **Decision Making Example 2**: Giving a treatment after a cancer test

$$l(y, a) = C_{y,a}$$

  - **Prediction Example**: Predicting the temperature in Potsdam
    - If we predict too high is as bad as too low and losses grow faster than the difference

$$l(y, \hat{y}) = (y - \hat{y})^2$$

		Actions	
		treat	nothing
Outcomes	Cancer	0	**1000**
	No cancer	**1**	0



Optimal Decisions: Expected Loss Minimization Principle

- **Expected Loss Minimization**. *Given a predictive model $p(y|x,D)$ and a loss function $l\colon \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$, the optimal action $a_{\mathrm{opt}}(x)$ is minimizing the expected loss*

$$a_{\mathrm{opt}}(x) \coloneqq \mathrm{argmin}_{a\in\mathcal{A}}\, E_{y\sim p(y|x,D)}[l(y,a)]$$

  - Optimal decisions require (yet again) solving an optimization problem!
  - **Decision Making Example**: Deciding which of n advertisement to show
    - Each expected loss is equal to $E_{y_a\sim p(y_a|x_a,D)}[-b_a \cdot y_a] = -P(y_a = 1|x,D) \cdot b_a$
    - The add with the largest expected bid $P(y_a = 1|x,D) \cdot b_a$ should be shown!
  - **Prediction Example**: *For squared loss* $y_{\mathrm{opt}} = E_{y\sim p(y|x,D)}[y]$
    - **Proof**: Taking the first derivative of $E_{y\sim p(y|x,D)}[l(y,\hat{y})]$ and setting it to zero gives

$$\frac{\mathrm{d}}{\mathrm{d}\hat{y}} E_{y\sim p(y|x,D)}[l(y,\hat{y})] = \sum_y p(y|x,D) \cdot \frac{\mathrm{d}}{\mathrm{d}\hat{y}} (y-\hat{y})^2$$

$$0 = \sum_y p(y|x,D) \cdot \left(2 \cdot \left(y - y_{\mathrm{opt}}\right)\right)$$

$$0 = 2 \cdot \left(\sum_y p(y|x,D) \cdot y - \sum_y p(y|x,D) \cdot y_{\mathrm{opt}}\right)$$

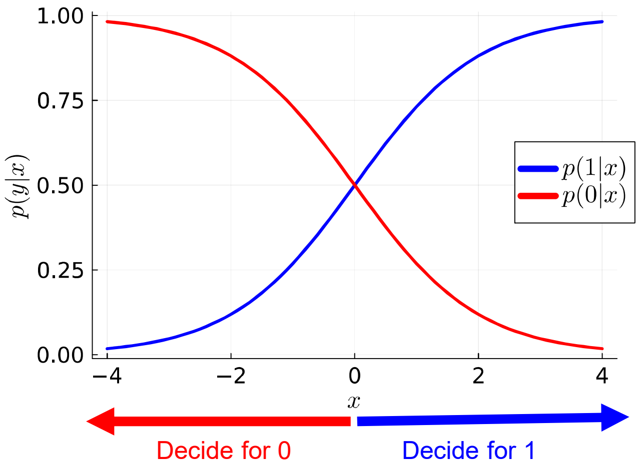
$$0 = E_{y\sim p(y|x)}[y] - y_{\mathrm{opt}}$$

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

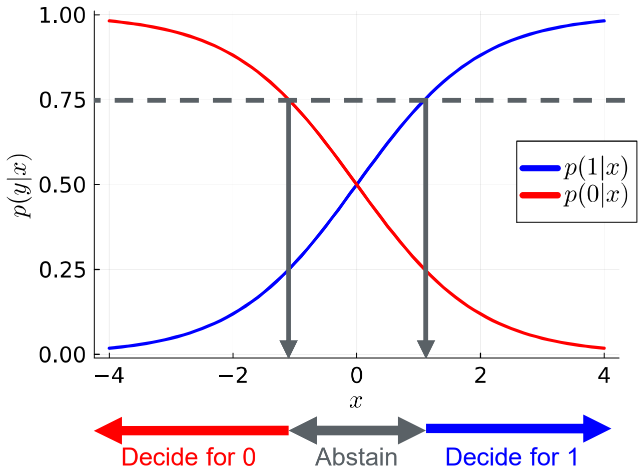
Decision Making under Uncertainty

- **Problem**: In practice, we rarely have predictive distributions $p(y|x, D)$ which are concentrated on a single outcome $\hat{y} \in \mathcal{Y}$.
- **Idea**: Introduce the action "to abstain" from making a prediction (and use human intelligence!)
  - **Example**: Automatically deciding on a treatment (health) or job application (business)

Optimal Binary Prediction (without abstaining)



Optimal Binary Prediction (with abstaining)



Summary

1. **Inference Methods**
   - Inference is the task of inferring what we know about the plausibility of a prediction function in light of training data
   - Bayesian Inference is the only consistent inference technique, but it requires huge summations which is (usually) computationally too hard
   - Maximum Likelihood Estimation is often easier and reduces machine learning to parameter optimization – but we are losing model uncertainty
2. **Decision Making**
   - Decision making solves the second big problem: making automated decisions based on future predictions
   - We *always* require domain-specific loss functions
   - Except in a few special cases (e.g., squared loss), automated decision making requires heavy optimization (again!)
   - We will not dive deeper into decision making methods in the rest of the lecture

See you next week!