CS5824: Advanced Machine Learning

Dawei Zhou

CS, Virginia Tech

Please keep your face covering on!

COMPUTER SCIENCE
VIRGINIA TECH

MLE
Linear Regression

Your first consulting job

- A billionaire asks you a question:
  - He says: I have a thumbtack, if I flip it, what's the probability it will fall with the nail up?
  - You say: Please flip it a few times:
  - You say: The probability is:
  - **He says: Why???**
  - You say: Because…

Thumbtack – Binomial Distribution

[0,1]

- P(Heads) = θ, P(Tails) = 1-θ

- Flips are i.i.d.:
  - Independent events
  - Identically distributed according to Binomial distribution

- Sequence D of α_H Heads and α_T Tails

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H} (1-\theta)^{\alpha_T}$$

$\alpha_H = 5$, $\alpha_T = 2$

{H, H, H, H, H, T, T}

Maximum Likelihood Estimation

- **Data:** Observed set *D* of α_H Heads and α_T Tails
- **Hypothesis:** Binomial distribution
- Learning θ is an optimization problem
  - – What's the objective function?
- ***MLE***: Choose θ that maximizes the probability of observed data:

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \quad P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \quad \ln P(\mathcal{D} \mid \theta)\end{aligned}$$

error $= \hat{\theta} - \theta$

Your First Learning Algorithm

$$\hat{\theta} = \arg\max_{\theta} \ln P(\mathcal{D} \mid \theta)$$
$$= \arg\max_{\theta} \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

α_H : # of H

α_T : # of T

$\theta \in$

$D = \{\underbrace{H \ldots H}_{5}, T, T\}$

- Set derivative to zero: $\boxed{\dfrac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = 0}$

$f = \ln x$

$\frac{d}{d\theta}\, \alpha_H \ln\theta + \alpha_T \ln(1-\theta) = 0$

$\frac{df}{dx} = \frac{1}{x}$

$\frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} = 0$

$\frac{\alpha_H}{\theta} = \frac{\alpha_T}{1-\theta}$

$\alpha_H(1-\theta) = \alpha_T \cdot \theta$

$\theta = \frac{\alpha_H}{\alpha_T + \alpha_H} = \frac{5}{7}$

How Many Flips Do I Need?

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

$\frac{1}{5}$

- Billionaire says: I flipped 3 heads and 2 tails.
- You say: θ = 3/5, I can prove it!

$\frac{1}{50}$

- He says: What if I flipped 30 heads and 20 tails?
- You say: Same answer, I can prove it!
- **He says: What's better?**
- You say: Humm… The more the merrier???
- He says: Is this why I am paying you the big bucks???

Simple Bound

$\frac{3}{7}$

- For $N = \alpha_H + \alpha_T$, and $\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$ $\frac{3}{5}$

total exp

- Let θ^* be the true parameter, for any $\varepsilon>0$:

$$P(|\hat{\theta} - \theta^*| \geq \epsilon) \leq 2e^{-2N\epsilon^2}$$

P(mis from H) = .001

MLE

hyper para

100 → 1

N ↑ → Big Data

PAC Learning

- PAC: Probably Approximately Correct
- Billionaire says: I want to know the thumbtack parameter θ, within ε = 0.1, with probability at least 1-δ = 0.95. How many flips?

$$P(|\,\hat{\theta} - \theta^*\,| \geq \epsilon) \;\leq\; 2e^{-2N\epsilon^2}$$

$N = 37.22$

0.95

$2e^{-2N\varepsilon^2} = 0.95$

N

$2e^{-2 \cdot N \cdot (0.1)} = 0.95$

What about prior

- Billionaire says: Wait, I know that the thumbtack is “close” to 50-50. What can you do for me now?
- **You say: I can learn it the Bayesian way…**

- Rather than estimating a single θ, we obtain a distribution over possible values of θ

Bayesian Learning

- Use Bayes rule:

Likelihood

prior

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

constant

[0, 1]

KDE, neural

- Or equivalently:

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$$

PC

Bayesian Learning for Thumbtack

$$P(\theta \mid \mathcal{D}) \ \propto \ P(\mathcal{D} \mid \theta) P(\theta)$$

- Likelihood function is simply Binomial:

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

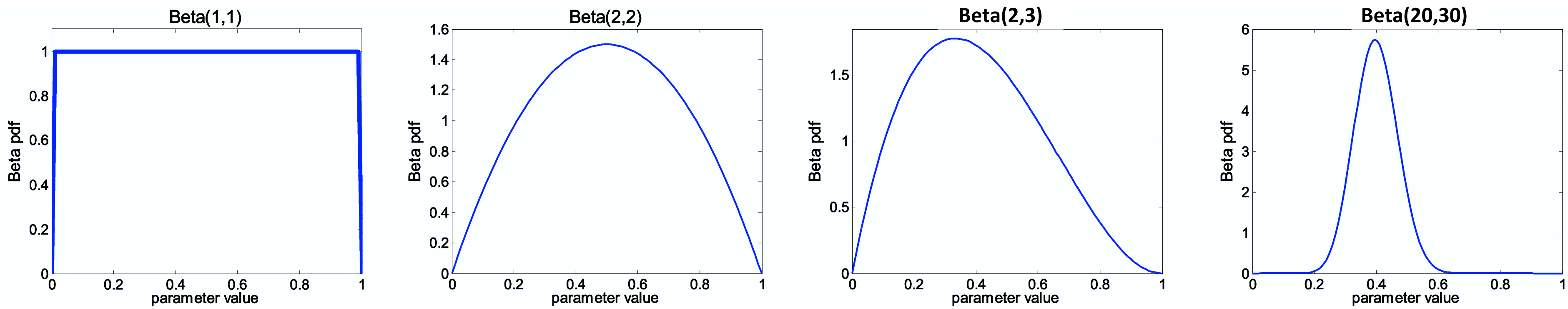
- What about prior?
  - Represent expert knowledge
  - Simple posterior form
- Conjugate priors:
  - Prior/posterior: same probability distribution family
  - **For Binomial, conjugate prior is Beta distribution**

Beta Prior Distribution – P(θ)

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

Mean: $\frac{b_H}{b_H + b_T}$

Mode: $\frac{b_H - 1}{b_H + b_T - 2}$



- Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$
- Posterior: $P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta) P(\theta)$

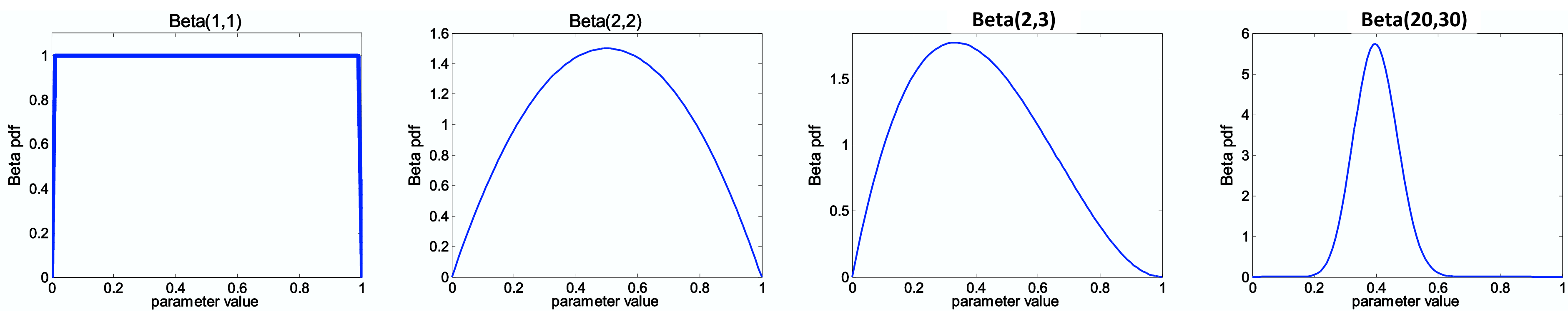
Posterior Distribution

$P(\theta|D) = P(D|\theta)\, P(\theta)$

- Prior: $Beta(\beta_H, \beta_T)$
- Data: α_H heads and α_T tails
- Posterior distribution:

Experiments

$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



real Exp: α_H, α_T
virtual Exp: β_H, β_T

Using Bayesian posterior

- Posterior distribution:

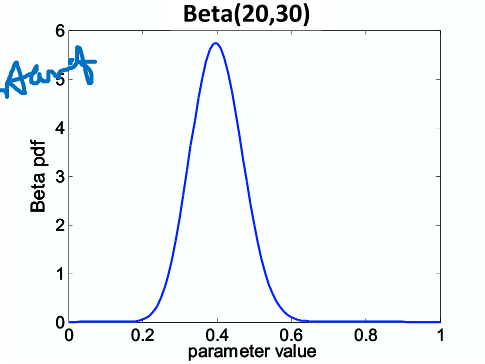
$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

- Bayesian inference:
  - No longer single parameter:

rv

$$E[f(\theta)] = \int_0^1 f(\theta) P(\underline{\theta} \mid \mathcal{D}) d\theta$$

constant



  - Integral is often hard to compute

MAP: Maximum a Posteriori Approximation

$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) d\theta$$

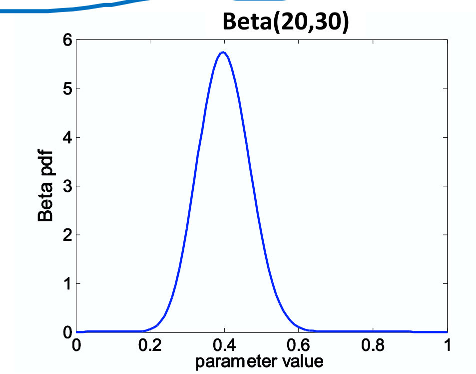
- As more data is observed, Beta is more certain

- MAP: use most likely parameter:

$$\hat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D}) \qquad E[f(\theta)] \approx f(\hat{\theta})$$

MAP for Beta Distribution

$$P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H+\alpha_H-1}(1-\theta)^{\beta_T+\alpha_T-1}}{B(\beta_H+\alpha_H,\beta_T+\alpha_T)} \sim Beta(\overset{20}{\beta_H}+\overset{\overset{10^6}{\uparrow}}{\overset{30}{\alpha_H}},\overset{20}{\beta_T}+\overset{\overset{10^7}{\uparrow}}{\overset{20}{\alpha_T}})$$



- MAP: use most likely parameter:

$$\widehat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D}) =$$

- Beta prior equivalent to extra thumbtack flips
- As $N \rightarrow \infty$, prior is "forgotten"
- **But, for small sample size, prior is important!**

What About Continuous Variables?

- Billionaire says: If I am measuring a continuous variable, what can you do for me?
- **You say: Let me tell you about Gaussians…**

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Some Properties of Gaussians

- Affine transformation (multiplying by scalar and adding a constant)
  - $X \sim N(\mu,\sigma^2)$
  - $Y = aX + b \rightarrow Y \sim N(a\mu+b,a^2\sigma^2)$

- Sum of Gaussians
  - $X \sim N(\mu_X,\sigma^2_X)$
  - $Y \sim N(\mu_Y,\sigma^2_Y)$
  - $Z = X+Y \rightarrow Z \sim N(\mu_X+\mu_Y, \sigma^2_X+\sigma^2_Y)$ Independence?

Learning a Gaussian

- Collect a bunch of data
  - Hopefully, i.i.d. samples
  - e.g., exam scores
- Learn parameters
  - Mean
  - Variance

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

MLE for Gaussian

- Prob. of i.i.d. samples *D*={x_1,…,x_N}:

$$P(\mathcal{D} \mid \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

- Log-likelihood of data:

$$\begin{aligned} \ln P(\mathcal{D} \mid \mu, \sigma) &= \ln\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}\right] \\ &= -N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} \end{aligned}$$

Your Second Learning Algorithm: MLE for Mean of a Gaussian

- What's MLE for mean?

$$\frac{d}{d\mu} \ln P(\mathcal{D} \mid \mu, \sigma) \;=\; \frac{d}{d\mu}\left[-N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

Properties of MLE for Mean

- Under certain conditions, MLE is consistent

$$\hat{m}_{MLE} \xrightarrow{P} m^*$$

- Asymptotic Normality: let $se = \sqrt{Var_m(\hat{m}_{MLE})}$.
  Under regularity conditions,

$$\frac{\widehat{\theta}_n - \theta}{se} \rightsquigarrow N(0,1) \qquad se \approx \sqrt{1/I_n(\theta)}$$

Fisher Information

MLE for Variance

- Again, set derivative to zero:

$$
\begin{aligned}
\frac{d}{d\sigma} \ln P(\mathcal{D} \mid \mu, \sigma) &= \frac{d}{d\sigma}\left[-N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}\right] \\
&= \frac{d}{d\sigma}\left[-N \ln \sigma\sqrt{2\pi}\right] - \sum_{i=1}^{N} \frac{d}{d\sigma}\left[\frac{(x_i - \mu)^2}{2\sigma^2}\right]
\end{aligned}
$$

Learning Gaussian Parameters

- MLE:

$$\widehat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\widehat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \widehat{\mu})^2$$

- BTW. MLE for the variance of a Gaussian is **biased**
  - Expected result of estimation is **not** true parameter!
  - Unbiased variance estimator:

$$\widehat{\sigma}^2_{unbiased} = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \widehat{\mu})^2$$

Bayesian Learning of Gaussian Parameters

- Conjugate priors
  - Mean: Gaussian prior
  - Variance: Wishart Distribution

- Prior for mean:

$$P(\mu \mid \eta, \lambda) = \frac{1}{\lambda\sqrt{2\pi}} e^{\frac{-(\mu-\eta)^2}{2\lambda^2}}$$

MAP for Mean of Gaussian

$$P(\mu \mid \eta, \lambda) = \frac{1}{\lambda\sqrt{2\pi}} e^{\frac{-(\mu-\eta)^2}{2\lambda^2}} \qquad P(\mathcal{D} \mid \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

$$\boxed{\frac{d}{d\mu}\left[\ln P(\mathcal{D} \mid \mu) P(\mu)\right] \;=\; \frac{d}{d\mu}\left[\ln P(\mathcal{D} \mid \mu) + \ln P(\mu)\right]}$$

Frequentist Statistics

- Data are random
- Estimators are random because they are functions of data
- Parameters are fixed, unknown constants not subject to probabilistic statements
- Procedures are subject to probabilistic statements, for example 95% confidence intervals trap the true parameter value 95% of the time
- Classifiers, even learned with deterministic procedures, are random because the training set is random
- PAC bound is frequentist

Bayesian Statistics

- Probability refers to degree of belief
- Inference about a parameter θ is by producing a probability distributions on it
- Starts with prior distribution p(θ)
- Likelihood function p(x | θ), a function of θ not x
- After observing data x, one applies the Bayes rule to obtain the posterior
- Prediction by integrating parameters out:

$$p(x \mid Data) = \int p(x \mid \theta) p(\theta \mid Data) d\theta$$

Prediction of Continuous Variables

- Billionaire says: Wait, that’s not what I meant!
- You says: Chill out, dude.
- He says: I want to predict a continuous variable for continuous inputs: I want to predict salaries from GPA.
- You say: **I can regress that…**

The Regression Problem

- **Instances:** <$\mathbf{x}_j$, t_j>
- **Learn:** Mapping from x to t(**x**)
- **Hypothesis space:**
  - Given, basis functions
  - Find coeffs **w**={w_1,…,w_k}

$$H = \{h_1, \ldots, h_K\}$$

$$\underbrace{t(\mathbf{x})}_{\text{data}} \approx \widehat{f}(\mathbf{x}) = \sum_i w_i h_i(\mathbf{x})$$

  - Why is this called linear regression???
    - model is linear in the parameters

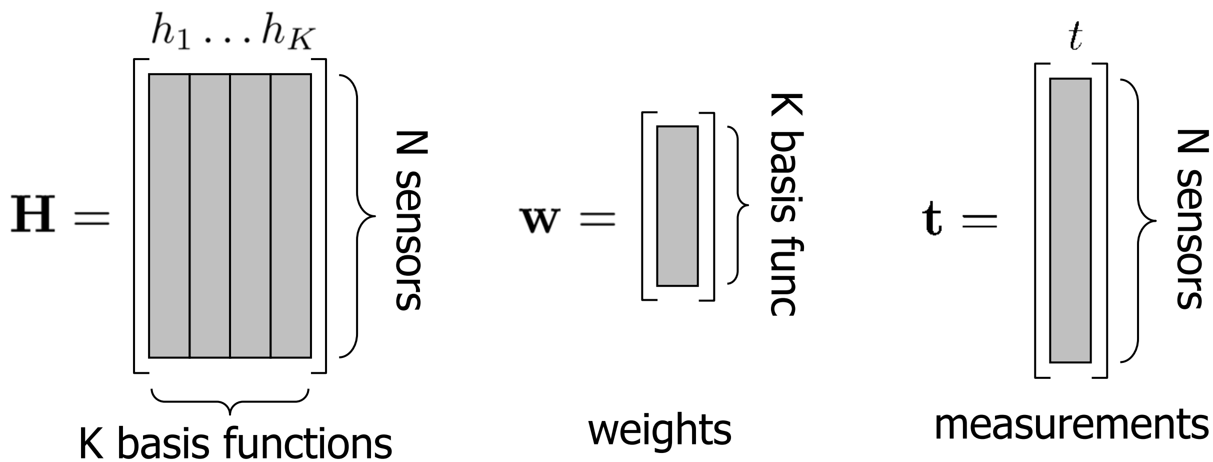
- **Precisely, minimize the residual squared error:**

$$\mathbf{w}^* \quad = \quad \arg\min_{\mathbf{w}} \sum_j \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

Regression in Matrix Notation

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_j \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

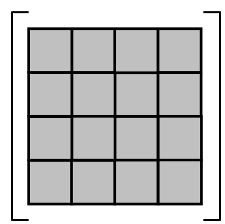
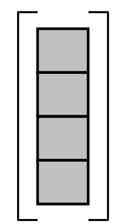
$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \underbrace{(\mathbf{H}\mathbf{w} - \mathbf{t})^T (\mathbf{H}\mathbf{w} - \mathbf{t})}_{\text{residual error}}$$



Regression Solution: Matrix Operations

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \underbrace{(\mathbf{H}\mathbf{w} - \mathbf{t})^T(\mathbf{H}\mathbf{w} - \mathbf{t})}_{\text{residual error}}$$

$$\text{solution: } \mathbf{w}^* = \underbrace{(\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}}_{\mathbf{A}^{-1}} \underbrace{\mathbf{H}^{\mathrm{T}}\mathbf{t}}_{\mathbf{b}} = \mathbf{A}^{-1}\mathbf{b}$$

where $\mathbf{A} = \mathbf{H}^{\mathrm{T}}\mathbf{H} =$  $\mathbf{b} = \mathbf{H}^{\mathrm{T}}\mathbf{t} =$ 

k×k matrix
for k basis functions

k×1 vector

But, Why?

- Billionaire (again) says: Why sum squared error???
- You say: Gaussians, Dr. Gateson, Gaussians…

- Model: prediction is linear function plus Gaussian noise
  - $t = \sum_i w_i h_i(\mathbf{x}) + \varepsilon$
- Learn **w** using MLE

$$P(t \mid \mathbf{x}, \mathbf{w}, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-\left[t - \sum_i w_i h_i(\mathbf{x})\right]^2}{2\sigma^2}}$$

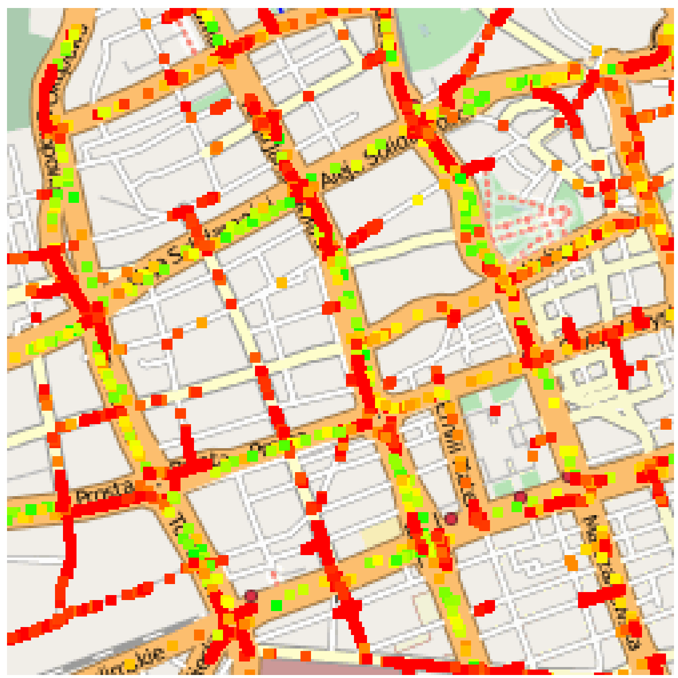
Least-squares Linear Regression is MLE for Gaussians!!!

Applications Corner 1

- Predict stock value over time from
  - past values
  - other relevant vars
    - e.g., weather, demands, etc.

Applications Corner 2

- Predict road traffic volume over time from
  - historical traffic volume
  - historical traffic volume of adjacent road segments



Applications Corner 3

- Predict when a sensor will fail
  - Based on several variables
    - age, chemical exposure, number of hours used,…

- ***Other applications?***