ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

inf | Informatik
Computer Science

Introduction to Machine Learning

A statistical perspective on supervised learning

Prof. Andreas Krause
Learning and Adaptive Systems (las.ethz.ch)

Motivation

- We have seen how we can fit prediction models (linear, non-linear) for regression and classification
- So far, these models do not have any statistical interpretation
- Often we would like to statistically model the data:
  - Quantify uncertainty
  - Express prior knowledge / assumptions about the data
- In the following, we will see how many of the approaches we have discussed can be interpreted as fitting probabilistic models
- This view will allow us to derive new methods

Recall: Goal of supervised learning

- Given training data

$$D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \subseteq \mathcal{X} \times \mathcal{Y}$$

- Want to identify a hypothesis $h : \mathcal{X} \to \mathcal{Y}$, e.g.,
  - Linear regression $h(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$
  - Kernel regression $h(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x})$
  - Neural network (single hidden layer) $h(\mathbf{x}) = \sum_{i=1}^{k} w_i' \varphi(\mathbf{w}_i^T \mathbf{x})$
- **Goal**: Want to minimize prediction error (risk)

Minimizing generalization error

- Fundamental assumption: Our data set is generated independently and identically distributed (iid)

$$(\mathbf{x}_i, y_i) \sim P(\mathbf{X}, Y)$$

- Would like to identify a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ that minimizes the prediction error (risk)

$$R(h) = \int P(\mathbf{x}, y)\ell(y; h(\mathbf{x}))d\mathbf{x}dy = \mathbb{E}_{\mathbf{x},y}[\ell(y; h(\mathbf{x}))]$$

- Defined in terms of a loss function

Least-squares regression

- In least-squares regression, risk is

$$R(h) = \mathbb{E}_{\mathbf{x},y}[(y - h(\mathbf{x}))^2]$$

- Suppose (unrealistically) we *knew* $P(\mathbf{X}, Y)$
- Which h minimizes the risk then?

$$\min_{h:\mathbb{R}^d \to \mathbb{R}} R(h) = \min_h \mathbb{E}_{x,y}\left[(y - h(x))^2\right] = \min_h \mathbb{E}_x\left[\mathbb{E}_y\left[(y - h(x))^2 \mid X = x\right]\right]$$

$$= \mathbb{E}_x\left[\min_{h(x)} \mathbb{E}_y\left[(y - h(x))^2 \mid X = x\right]\right]$$

since we consider arbitrary h, choose $h(x)$ and $h(x')$ independently for $x = x'$

Least-squares regression

For a given x, what is the optimal prediction?

$$y^*(x) \in \operatorname*{argmin}_{\hat{y}} \underbrace{\mathbb{E}_y\left[(\hat{y}-Y)^2 \mid X=x\right]}_{\ell(\hat{y})}$$

$$\ell(\hat{y}) = \int (\hat{y}-y)^2 p(y|x)\, dy$$

$$\frac{d}{d\hat{y}}\ell(\hat{y}) = \int \frac{d}{d\hat{y}}(\hat{y}-y)^2 p(y|x)\, dy = \int 2(\hat{y}-y)\, p(y|x)\, dy \stackrel{!}{=} 0$$

$$\Rightarrow \underbrace{\int \hat{y}\, p(y|x)\, dy}_{\hat{y}} = \underbrace{\int y\, p(y|x)\, dy}_{\mathbb{E}[Y|X=x]} \quad \Leftrightarrow \quad \hat{y} = \mathbb{E}[Y|X=x]$$

Minimizing the least squares error

- Assuming the data is generated iid according to

$$(\mathbf{x}_i, y_i) \sim P(\mathbf{X}, Y)$$

- The hypothesis h* minimizing $R(h) = \mathbb{E}_{\mathbf{x},y}[(y - h(\mathbf{x}))^2]$ is given by the conditional mean

$$h^*(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$$

- This (in practice unattainable) hypothesis is called the Bayes‘ optimal predictor for the squared loss

In practice we have finite data

- We know that

$$h^*(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$$

- Thus, one strategy for estimating a predictor from training data is to estimate the conditional distribution

$$\hat{P}(Y \mid \mathbf{X})$$

and then, for test point ***x***, predict label

$$\hat{y} = \hat{\mathbb{E}}[Y \mid \mathbf{X} = \mathbf{x}] = \int \hat{P}(y \mid \mathbf{X} = \mathbf{x})\, y\, dy$$

Estimating conditional distributions

- Common approach: Parametric estimation
  - Choose a particular parametric form $\hat{P}(Y \mid \mathbf{X}, \theta)$
  - Then optimize the parameters. How?

➔ Maximum (conditional) Likelihood Estimation

$$\theta^* = \arg\max_{\theta} \hat{P}(y_1, \ldots, y_n \mid \mathbf{x}_1, \ldots, \mathbf{x}_n, \theta)$$

$$\overset{iid}{=} \arg\max_{\theta} \prod_{i=1}^{n} \hat{P}(y_i \mid x_i, \theta) = \arg\max_{\theta} \log \prod_{i=1}^{n} \hat{P}(y_i \mid x_i, \theta)$$

$$= \arg\max_{\theta} \sum_{i=1}^{n} \log \hat{P}(y_i \mid x_i, \theta) = \arg\min_{\theta} -\sum_{i=1}^{n} \log \hat{P}(y_i \mid x_i, \theta)$$

Example: Conditional linear Gaussian

Sps. assume $Y = h(X) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$

assume $h(x) = w^T x$

$\hookrightarrow \hat{P}(Y | X, w, \sigma^2) = N(y; w^T x, \sigma^2)$ ← known

$\hookrightarrow \hat{w} = \underset{w}{\operatorname{argmax}} \hat{P}(y_{1:n} | x_{1:n}, w, \sigma^2)$

$= \underset{w}{\operatorname{argmin}} - \sum_{i=1}^{n} \log \hat{P}(y_i | x_i, w, \sigma^2)$

A probabilistic model for regression

- Consider linear regression. Let's make the statistical assumption that the noise is Gaussian:

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

- Then we can compute the (conditional) likelihood of the data given any candidate model ***w*** as:

$$-\log \hat{p}(y|x,w) = -\log \mathcal{N}(y|w^T x, \sigma^2) = -\log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-w^T x)^2}{2\sigma^2}\right)$$

$$= \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}(y - w^T x)^2$$

$$\Rightarrow \hat{w} = \arg\max_w P(y_{1:n}|x_{1:n}, w, \sigma^2) = \arg\min_w \sum_{i=1}^n \left(\frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}(y_i - w^T x_i)^2\right)$$

$$= \arg\min \underbrace{\frac{n}{2}\log 2\pi\sigma^2}_{\text{const. w.r.t. } w} + \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - w^T x_i)^2 = \arg\min_w \sum_{i=1}^n (y_i - w^T x_i)^2$$

MLE for conditional linear Gaussian

- The negative log likelihood is given by

$$L(\mathbf{w}) = -\log P(y_1, \ldots, y_n \mid \mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{w}) = \frac{n}{2}\log(2\pi\sigma^2) + \sum_{i=1}^{n} \frac{(y_i - \mathbf{w}^T\mathbf{x}_i)^2}{2\sigma^2}$$

- Thus, under the „conditional linear Gaussian“ assumption, maximizing the likelihood is equivalent to least squares estimation:

$$\arg\max_{\mathbf{w}} P(y_1, \ldots, y_n \mid \mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{w}) = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - \mathbf{w}^T\mathbf{x}_i)^2$$

More generally: MLE for iid Gaussian noise

- Suppose $\mathcal{H} = \{h : \mathcal{X} \to \mathbb{R}\}$ is a class of functions

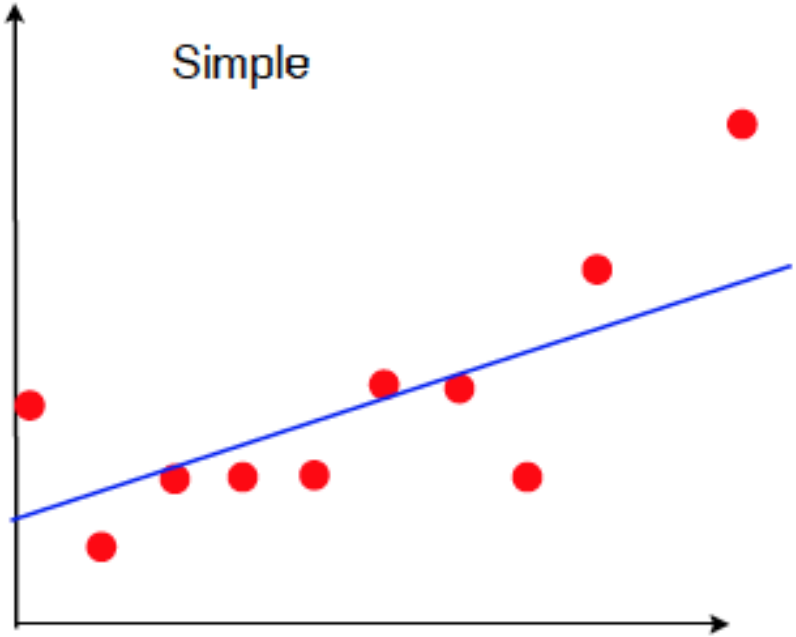
- Assuming that $P(Y = y \mid \mathbf{X} = \mathbf{x}) = \mathcal{N}(y \mid h^*(\mathbf{x}), \sigma^2)$ for some function $h^* : \mathcal{X} \to \mathbb{R}$ and some $\sigma^2 > 0$ the MLE for data $D = \{(\mathbf{x}_1, y_1), \ldots (\mathbf{x}_n, y_n)\}$ in $\mathcal{H}$ is given by

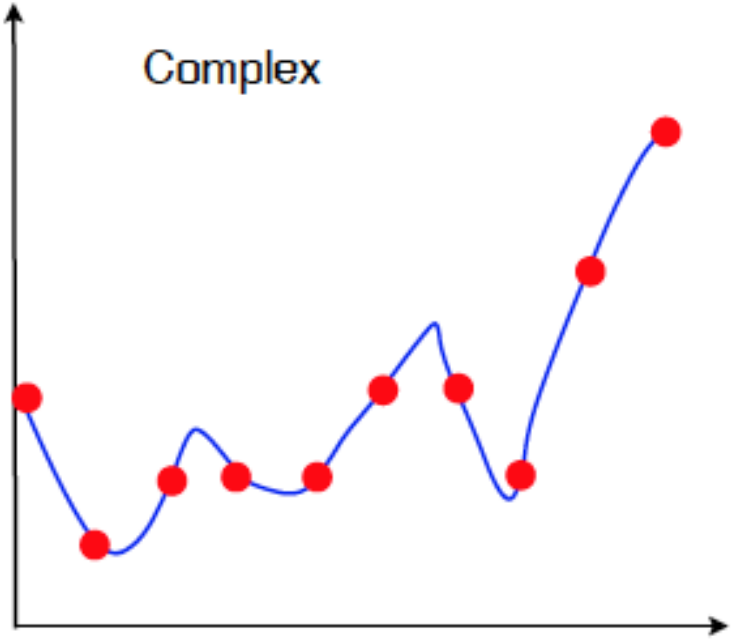
$$\hat{h} = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{n} (y_i - h(\mathbf{x}_i))^2$$

Least-squares regression = Gaussian MLE

- The **Maximum Likelihood Estimate (MLE) is given by the least squares solution**, assuming that the noise is iid Gaussian with constant variance
- This is useful since MLE satisfies several nice statistical properties (not formally defined here)
  - Consistency (parameter estimate converges to true parameters in probability)
  - Asymptotic efficiency (smallest variance among all „well-behaved“ estimators for large n)
  - Asymptotic normality
- However, all these properties are asymptotic (hold as $n \to \infty$). For finite n, we must avoid overfitting!

Recall: Overfitting in regression





Bias Variance Tradeoff

$$\textit{Prediction error} = \textit{Bias}^2 + \textit{Variance} + \textit{Noise}$$

- **Bias**: Excess risk of best model *considered* compared to minimal achievable risk knowing *P(X,Y)* (i.e., given infinite data)
- **Variance**: Risk incurred due to estimating model from limited data
- **Noise**: Risk error incurred by optimal model (i.e., irreducible error)

Bias in estimation

- MLE solution depends on training data *D*

$$\hat{h} = \hat{h}_D = \arg\min_{h \in \mathcal{H}} \sum_{(\mathbf{x},y) \in D} (y - h(\mathbf{x}))^2$$

- But training data *D* is itself random (drawn iid from P)
- We might want to choose *H* to have small bias (i.e., have small squared error on average)

$$\mathbb{E}_X \left[\mathbb{E}_D \hat{h}_D(\mathbf{X}) - h^*(\mathbf{X}) \right]^2$$

Bias

Variance in estimation

- MLE solution depends on training data *D*

$$\hat{h} = \hat{h}_D = \arg\min_{h \in \mathcal{H}} \sum_{(\mathbf{x},y) \in D} (y - h(\mathbf{x}))^2$$

- This estimator is itself random, and has some variance

$$\mathbb{E}_{\mathbf{X}} \mathrm{Var}_D \left[\hat{h}_D(\mathbf{X}) \right]^2$$

$$= \mathbb{E}_{\mathbf{X}} \mathbb{E}_D \left[\hat{h}_D(\mathbf{X}) - \mathbb{E}_{D'} \hat{h}_{D'}(\mathbf{X}) \right]^2$$

Noise in estimation

- Even if we know the Bayes‘ optimal hypothesis h*, we‘d still incur some error due to noise

$$\mathbb{E}_{\mathbf{X},Y}[(Y - h^*(\mathbf{X}))^2]$$

- This error is irreducible, i.e., independent of choice of the hypothesis class

Bias-variance tradeoff

- For least-squares estimation the following holds

Expected risk

$$\mathbb{E}_D \mathbb{E}_{\mathbf{X},Y}[(Y - \hat{h}_D(\mathbf{X}))^2]$$

$$= \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_D \hat{h}_D(\mathbf{X}) - h^*(\mathbf{X})\right]^2 \quad \text{Bias}^2$$

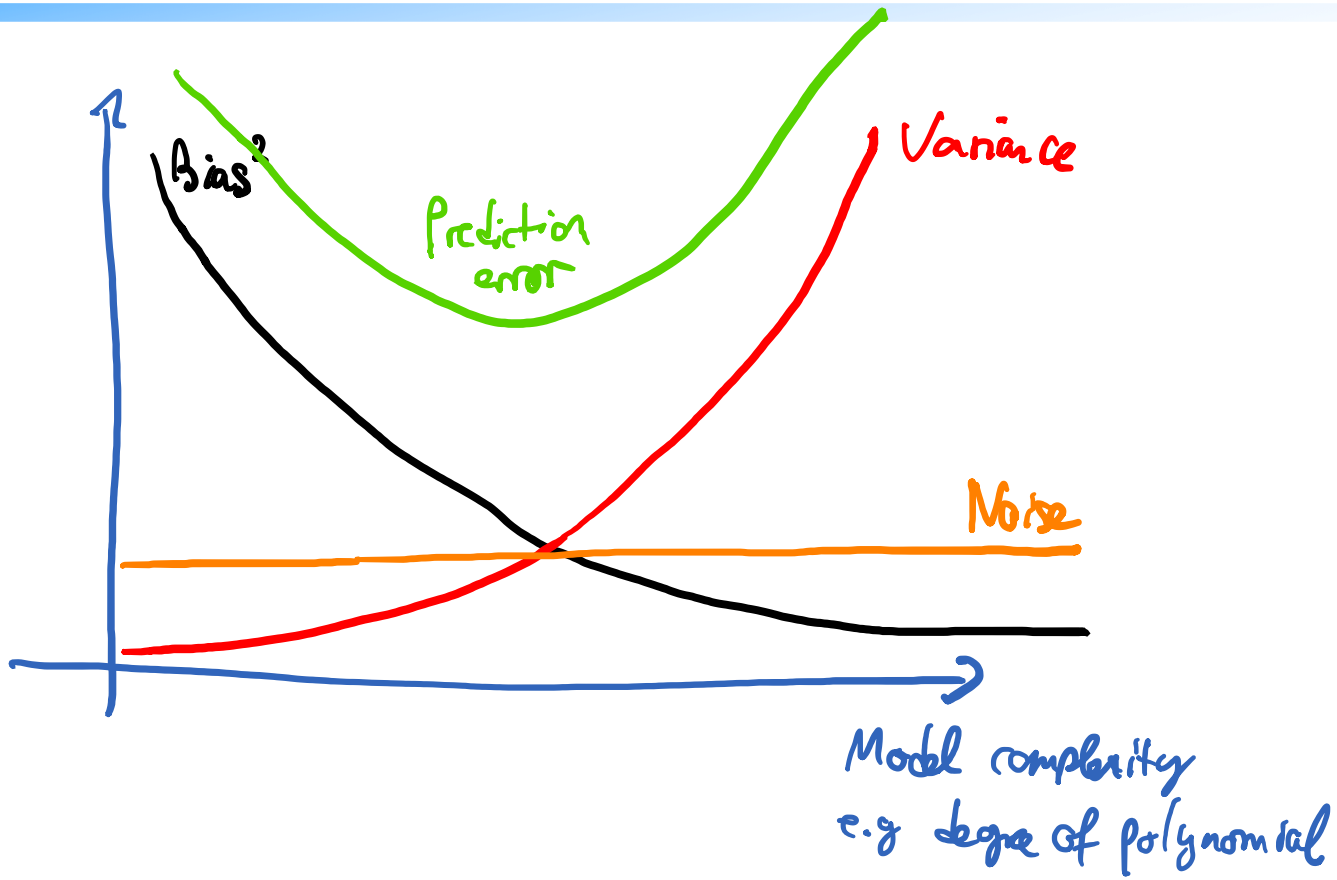
Bias

$$+\mathbb{E}_{\mathbf{X}}\mathbb{E}_D\left[\hat{h}_D(\mathbf{X}) - \mathbb{E}_{D'}\hat{h}_{D'}(\mathbf{X})\right]^2 \quad \text{Variance}$$

$$+\mathbb{E}_{\mathbf{X},Y}[Y - h^*(\mathbf{X})]^2 \quad \text{Noise}$$

➔ Ideally wish to find estimator that simultaneously minimizes bias and variance

Bias variance tradeoff illustration



Bias-variance demo

Bias and variance in regression

- The maximum likelihood estimate (= least-squares fit) for linear regression is unbiased (if h^* in class H)
- Furthermore, it is the minimum variance estimator among all unbiased estimators
  (Gauss-Markov Theorem, not explained further here)
- However, we have already seen that the least-squares solution can overfit

➔ Thus, trade (a little bit of) bias for a (potentially dramatic) reduction in variance

➔ Regularization (e.g., ridge regression, Lasso, ...)

Summary: Bias Variance Tradeoff

$$\textit{Prediction error} = \textit{Bias}^2 + \textit{Variance} + \textit{Noise}$$

- **Bias**: Excess risk of best model *considered* compared to minimal achievable risk knowing *P(X,Y)* (i.e., given infinite data)
- **Variance**: Risk incurred due to estimating model from limited data
- **Noise**: Risk error incurred by optimal model (i.e., irreducible error)

Trade bias and variance via model selection / regularization

Introducing bias through Bayesian modeling

- Can introduce bias by expressing assumptions on parameters through a Bayesian prior
- For example, let's assume $\mathbf{w} \sim \mathcal{N}(0, \beta^2 \mathbf{I})$

$w_i \sim \mathcal{N}(0, \beta^2)$

- Then, the posterior distribution of ***w*** is given using Bayes' rule by

$$P(w \mid x_{1:n}, y_{1:n}) = \frac{P(w \mid x_{1:n})\, P(y_{1:n} \mid w, x_{1:n})}{P(y_{1:n} \mid x_{1:n})}$$

Apply Bayes' rule to conditional distributions (cond. on $x_{1:n}$)

$$= \frac{P(w)\, P(y_{1:n} \mid w, x_{1:n})}{P(y_{1:n} \mid x_{1:n})}$$

assumes w indep. of x

- Which parameters ***w*** are most likely a posteriori?

Maximum a posteriori estimate

$$P(\mathbf{w} \mid \mathbf{x}_1, \ldots, \mathbf{x}_n, y_1, \ldots, y_n) = \frac{P(\mathbf{w})P(y_1, \ldots, y_n \mid \mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{w})}{P(y_1, \ldots, y_n \mid \mathbf{x}_1, \ldots, \mathbf{x}_n)}$$

(*) $\operatorname{argmax}_w P(w \mid x_{1:n}, y_{1:n}) = \operatorname{argmin}_w -\log P(w) - \log P(y_{1:n} \mid x_{1:n}, w) + \log P(y_{1:n} \mid x_{1:n})$

indep. of w

$$-\log P(w) = -\log \prod_{i=1}^{d} P(w_i) = -\sum_{i=1}^{d} \log \mathcal{N}(w_i; 0, \beta^2)$$

$$= -\sum_{i=1}^{d} \log \frac{1}{\sqrt{2\pi\beta^2}} \exp(-w_i^2/2\beta^2) = \frac{d}{2}\log 2\pi\beta^2 + \frac{1}{\beta^2}\sum_{i=1}^{d} w_i^2$$

$$= \text{Const} + \frac{1}{2\beta^2}\|w\|_2^2$$

$$(*) = \operatorname{argmin}_w \frac{1}{2\beta^2}\|w\|_2^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - w^T x_i)^2$$

$$= \operatorname{argmin}_w \underbrace{\frac{\sigma^2}{\beta^2}}_{\lambda}\|w\|_2^2 + \sum_{i=1}^{n}(y_i - w^T x_i)^2 \quad \Rightarrow \text{Ridge regression for } \lambda = \frac{\sigma^2}{\beta^2}$$

Ridge regression = MAP estimation

- Ridge regression can be understood as finding the Maximum A Posteriori (MAP) parameter estimate for a linear regression problem, assuming that
  - The noise *P(y|**x**,**w**)* is iid Gaussian and
  - The prior *P(**w**)* on the model parameters ***w*** is Gaussian

$$\arg\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda ||\mathbf{w}||_2^2 \equiv \arg\max_{\mathbf{w}} P(\mathbf{w}) \prod_i P(y_i \mid \mathbf{x}_i, \mathbf{w})$$

Regularization vs. MAP inference

- More generally, regularized estimation can often be understood as MAP inference

$$\arg\min_{\mathbf{w}} \sum_{i=1}^{n} \ell(\mathbf{w}^T\mathbf{x}_i; \mathbf{x}_i, y_i) + C(\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_i P(y_i \mid \mathbf{x}_i, \mathbf{w}) P(\mathbf{w})$$

$$= \arg\max_{\mathbf{w}} P(\mathbf{w} \mid D)$$

where $C(\mathbf{w}) = -\log P(\mathbf{w})$

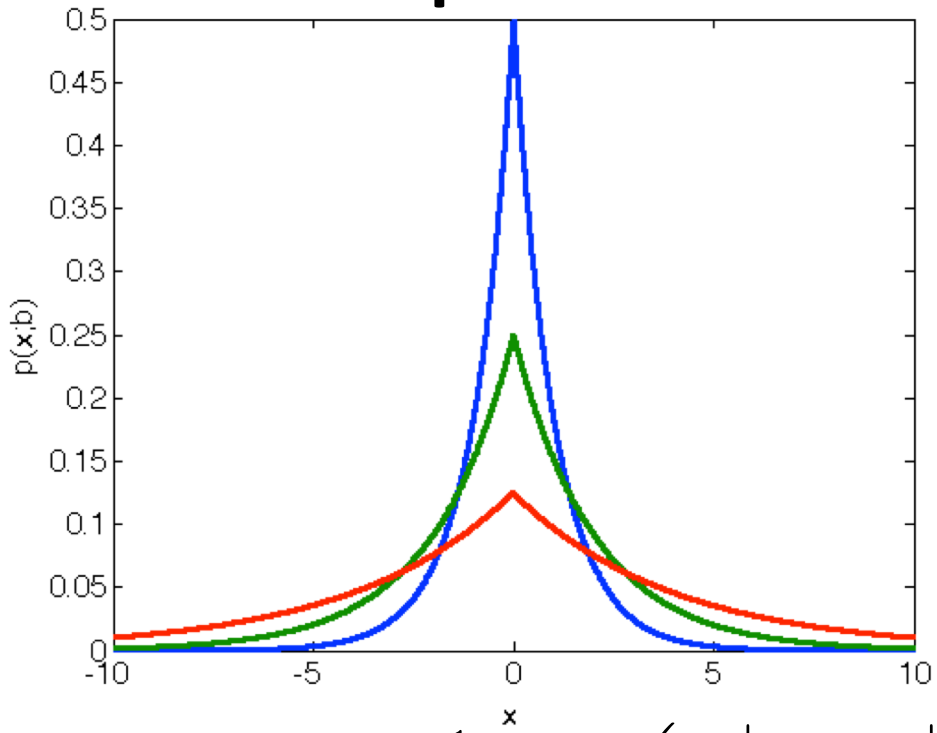
and $\ell(\mathbf{w}^T\mathbf{x}_i; \mathbf{x}_i, y_i) = -\log P(y_i \mid \mathbf{x}_i, \mathbf{w})$

- This perspective allows changing priors (=regularizers) and likelihoods (=loss functions)

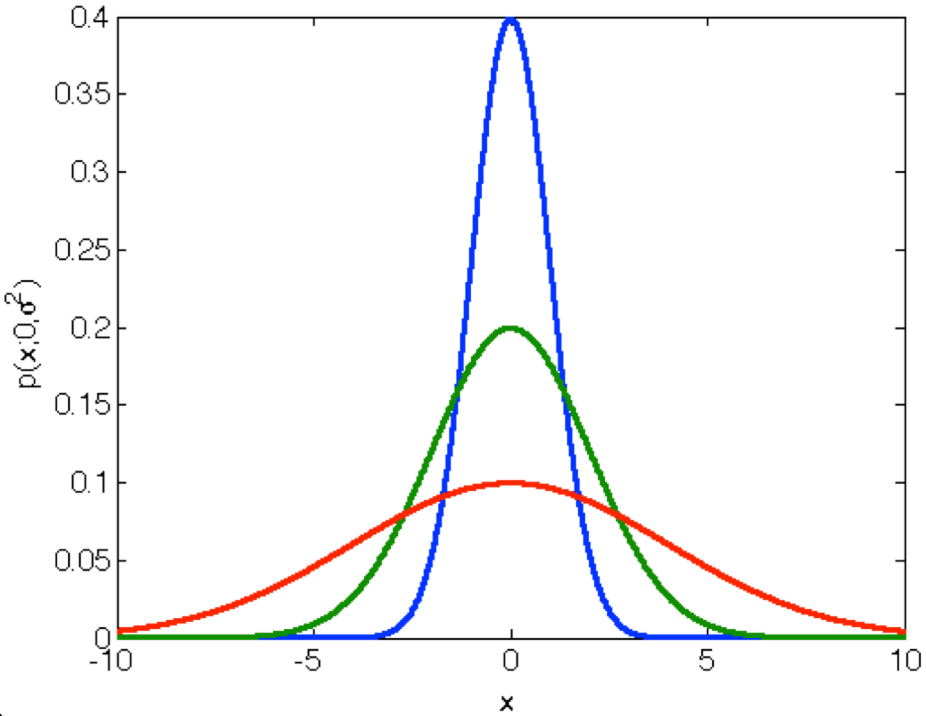
Example: l1-regularization

- Is there a prior that corresponds to l1-regularization?
- **Answer**: The Laplace prior

Laplace



Compare with **Gaussian**

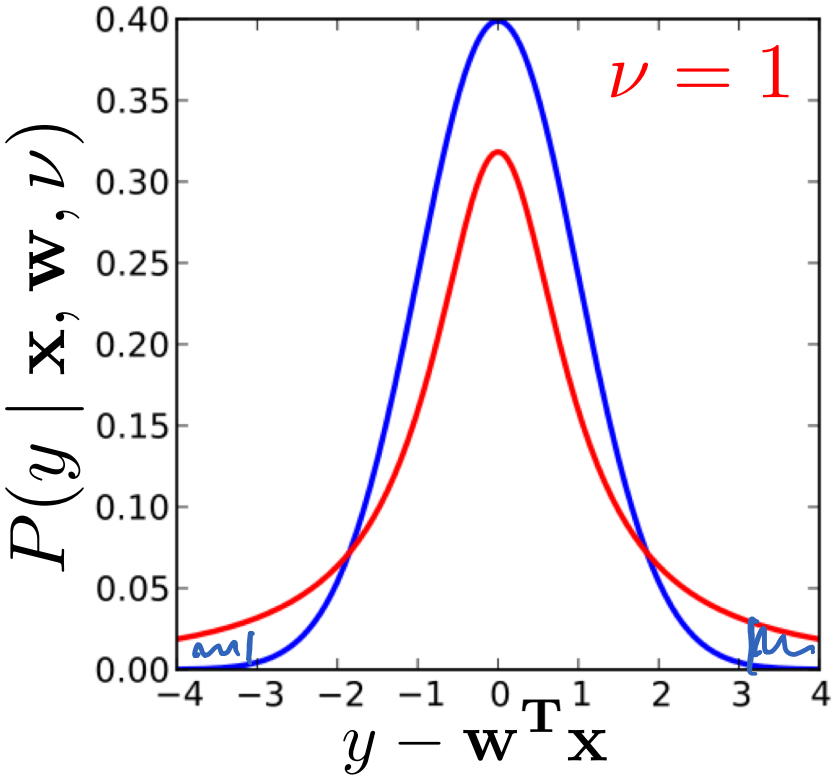


$$p(x;\mu,b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$

Example: student-t likelihood

- Can introduce robustness by changing the likelihood (=loss) function
- **Example**: (non-standardized) Student's-t likelihood

$$P(y \mid \mathbf{x}, \mathbf{w}, \nu, \sigma^2) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu\sigma^2}\Gamma(\frac{\nu}{2})}\left(1 + \frac{(y - \mathbf{w}^T\mathbf{x})^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}$$



For Gaussian: $P(|y - w^T x| > t \cdot \sigma^2) = O(\exp -t)$

For student-t: $\sim$ " —— $= O(t^{-\nu})$