Chapter 8: Where Models Meet Data

Mathematics for Machine Learning

San Diego Machine Learning
Liam Barstad

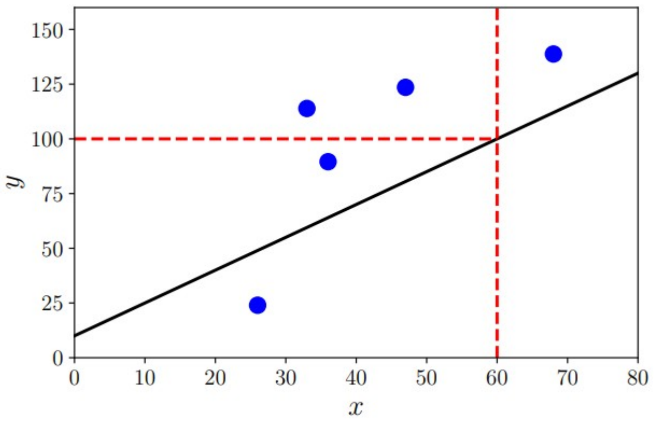
Models

Models as Functions (Predictor)

Deterministic output

$$f : \mathbb{R}^D \to \mathbb{R}$$

$$f(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{x} + \theta_0$$

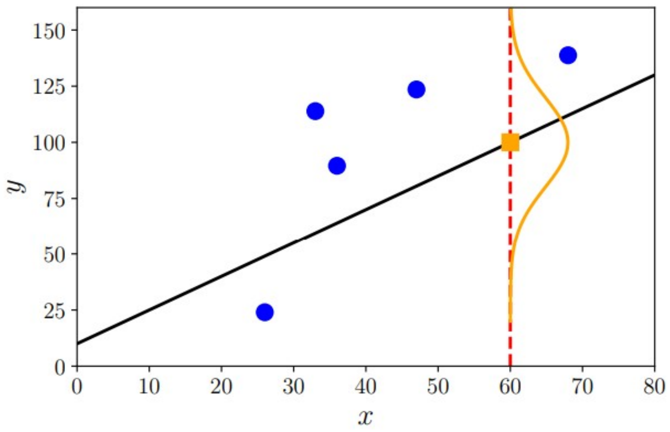


Models as Distributions

Quantifies uncertainty

Allows for noisy data

Parameters = Statistics



Examples: Logistic Regression, Naive Bayes, Gaussian Processes, Hidden Markov Models (HMMs)

Hyperparameters set before training to influence model structure and behavior

- Learning rate
- Number of layers
- Batch size

Empirical Risk Minimization

$$\mathbf{R}_{\text{emp}}(f, \boldsymbol{X}, \boldsymbol{y}) = \frac{1}{N} \sum_{n=1}^{N} \ell(y_n, \hat{y}_n)$$

Sample - **(Empirical Risk)**

- Depends on model *f*, data **X**, and labels *y*
- Assumes data points are IID

$$\mathbf{R}_{\text{true}}(f) = \mathbb{E}_{\boldsymbol{x},y}[\ell(y, f(\boldsymbol{x}))]$$

Population - **(Expected Risk)**

Overfitting – R_{emp} underestimates R_{true} – little data for complex hypotheses – possibly too many parameters – all modeling power used to reduce training error

Underfitting – R_{emp} and R_{true} are high – model too simple

Regularization – Penalizes overly flexible predictors, makes model better at generalizing

$$\min_{\boldsymbol{\theta}} \frac{1}{N} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2 \qquad \min_{\boldsymbol{\theta}} \frac{1}{N} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2 + \lambda \|\boldsymbol{\theta}\|^2$$

(least squares, Tikhonov/L2/Ridge regression)

(Alternative) Ivanov Regularization

Constrains the regularization parameter to be less than some value

$$h: \quad \arg\min_{\boldsymbol{w},b} \quad \hat{L}(h)$$
$$s.t. \quad \|\boldsymbol{w}\|^2 \leq w^2_{\mathrm{MAX}},$$

ChatGPT says you can also add smoothing function to regularization parameter (Image processing?)

$$\min_{\mathbf{w}} \left(\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda \|\mathbf{L}\mathbf{w}\|^2 \right)$$

Morozov Regression: constrains loss

$$\min_{x \in \mathrm{X}} \mathcal{R}(x) \quad \text{s.t.} \quad \|\mathbf{A}x - y_\delta\| \leq \delta\,.$$

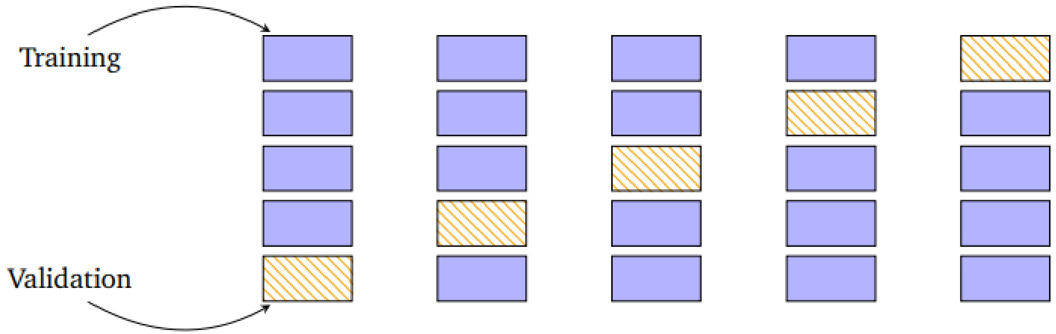
Tikhonov, Ivanov, Morozov regression for SVMs: https://link.springer.com/article/10.1007/s10994-015-5540-x
(There exists a regularization parameter such that all 3 are equivalent)

Morozov regression example: https://arxiv.org/pdf/2310.14290

Cross-Validation

Validation set – subset of data kept aside to evaluate performance of model

K-Fold Cross-Validation – embarassingly parallel



$$\mathbb{E}_{\mathcal{V}}[R(f, \mathcal{V})] \approx \frac{1}{K} \sum_{k=1}^{K} R(f^{(k)}, \mathcal{V}^{(k)})$$

(Alternative) Bootstrap and Jackknife

Bootstrap – Method for estimating distributions, sampling with replacement

- Estimates confidence intervals, precision, standard error
- Calculates sample sizes
- Deals with non-normal data

Jackknife – Sequentially deletes samples one by one, then recomputing statistic

Jackknife After Bootstrap – determines how well sample created by bootstrapping represents population

Jackknife/Bootstrap Overview:
https://www.datasciencecentral.com/resampling-methods-comparison

Original Paper:
https://www.math.wustl.edu/~kuffner/AlastairYoung/Efron1992discussion.pdf

Maximum Likelihood Estimation (MLE)

Negative Log Likelihood – probability of y_n given x_n, with parameters

$$\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{\theta}) = -\log p(\boldsymbol{x} \mid \boldsymbol{\theta})$$

$$p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}) = \mathcal{N}\left(y_n \mid \boldsymbol{x}_n^\top \boldsymbol{\theta}, \sigma^2\right)$$ (Gaussian)

$$p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) = \prod_{n=1}^{N} p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})$$ (IID)

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= -\sum_{n=1}^{N} \log p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}) = -\sum_{n=1}^{N} \log \mathcal{N}\left(y_n \mid \boldsymbol{x}_n^\top \boldsymbol{\theta}, \sigma^2\right) \\
&= -\sum_{n=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_n - \boldsymbol{x}_n^\top \boldsymbol{\theta})^2}{2\sigma^2}\right) \\
&= -\sum_{n=1}^{N} \log \exp\left(-\frac{(y_n - \boldsymbol{x}_n^\top \boldsymbol{\theta})^2}{2\sigma^2}\right) - \sum_{n=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \\
&= \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n^\top \boldsymbol{\theta})^2 - \sum_{n=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}}\,.
\end{aligned}$$

$$\mathcal{L}(\boldsymbol{\theta}) = -\log p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) = -\sum_{n=1}^{N} \log p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})$$

Turning the product into a sum makes it simpler and more numerically stable

Maximum A Posteriori Estimation (MAP)

Instead of estimating minimum of negative log **likelihood**, can measure minimum of negative log **posterior**

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\boldsymbol{x})} \qquad p(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$$

- Incorporates prior knowledge of parameter distribution through a conjugate prior (e.g. Gaussian), “where good parameters lie”
- Can act like regularization

MLE Properties:
- Converges the true value, plus an error that is normal, the error’s variance decaying in 1/N
- With small data, MLE can lead to overfitting

Prob. Models and Bayesian Inference

Probabilistic models have a consistent set of rules from probability theory

$$p(\boldsymbol{\theta} \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathcal{X})}, \qquad p(\mathcal{X}) = \int p(\mathcal{X} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta},$$

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = \mathbb{E}_{\boldsymbol{\theta}}[p(\boldsymbol{x} \mid \boldsymbol{\theta})],$$

Example: use a Bernoulli distribution (likelihood) with a Beta distribution prior to calculate how likely it is for X number of clicks to happen in a 5 minute window

- Topic modeling
- Click-through rate prediction
- Online ranking systems
- Large-scale recommender systems

Latent Variables

Defines process that generates data from parameters

E.g. observed data like heart rate and pupil dilation are related to the latent variable anxiety, and the even more latent variable procrastination

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{z})$$

To compute:

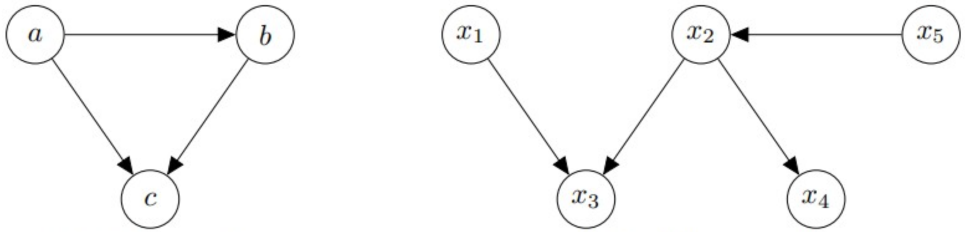
1. compute p(x|θ) without z
2. use likelihood for parameter estimation

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \int p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{z}) p(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \qquad p(\boldsymbol{z} \mid \mathcal{X}, \boldsymbol{\theta}) = \frac{p(\mathcal{X} \mid \boldsymbol{z}, \boldsymbol{\theta}) p(\boldsymbol{z})}{p(\mathcal{X} \mid \boldsymbol{\theta})},$$

$$p(\boldsymbol{z} \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \boldsymbol{z}) p(\boldsymbol{z})}{p(\mathcal{X})}, \qquad p(\mathcal{X} \mid \boldsymbol{z}) = \int p(\mathcal{X} \mid \boldsymbol{z}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta},$$

Directed Graphical Models

If arrow connects a to b, gives the probability p(b | a)



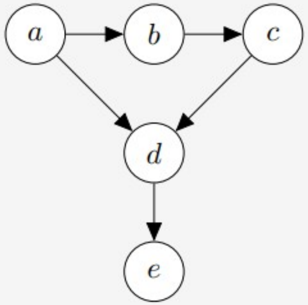
(a) Fully connected. (b) Not fully connected.

$$p(\boldsymbol{x}) = \prod_{k=1}^{K} p(x_k \mid \mathrm{Pa}_k)$$

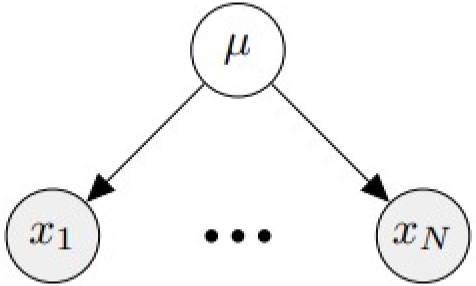
$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_5)p(x_2 \mid x_5)p(x_3 \mid x_1, x_2)p(x_4 \mid x_2)$$

D-separated if:

- Arrows meet head to tail or tail to tail, and node is in C
- Arrows meet head to head and no node or descendants meet in C



$$b \perp\!\!\!\perp d \mid a, c$$
$$a \perp\!\!\!\perp c \mid b$$
$$b \not\perp\!\!\!\perp d \mid c$$
$$a \not\perp\!\!\!\perp c \mid b, e$$



Probabilistic Model Example - HMM

Markov assumption: future state of a system only depends on its present state, not on past states

Assume that data is modeled by series of latent hidden states

- Transition probabilities: moving from one state to another
- Emission probabilities: observing an output given a hidden state

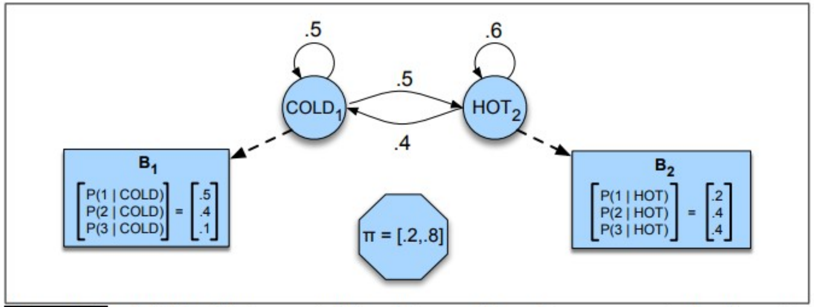
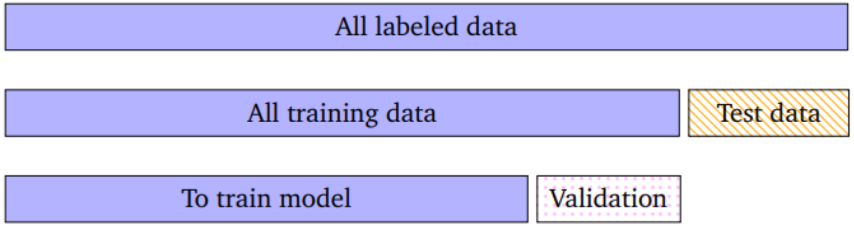


Figure A.2 A hidden Markov model for relating numbers of ice creams eaten by Jason (the observations) to the weather (H or C, the hidden variables).

Practical Example: https://www.geeksforgeeks.org/hidden-markov-model-in-machine-learning/
In-Depth Explanation: https://web.stanford.edu/~jurafsky/slp3/A.pdf

Model Selection

Nested Cross-Validation – Inner training loop as well as outer, inner loop is validation set for hyperparameter tuning, outer loop is test set



$$\mathbb{E}_{\mathcal{V}}[\mathbf{R}(\mathcal{V} \mid M)] \approx \frac{1}{K} \sum_{k=1}^{K} \mathbf{R}(\mathcal{V}^{(k)} \mid M)$$

Bayesian Model Selection - Instead of penalizing complex hypotheses through regularization, place prior on models

$$M_k \sim p(M)$$
$$\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta} \mid M_k)$$
$$\mathcal{D} \sim p(\mathcal{D} \mid \boldsymbol{\theta}_k)$$

$$p(M_k \mid \mathcal{D}) \propto p(M_k) p(\mathcal{D} \mid M_k)$$

$$p(\mathcal{D} \mid M_k) = \int p(\mathcal{D} \mid \boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k \mid M_k) d\boldsymbol{\theta}_k$$

$$M^* = \arg\max_{M_k} p(M_k \mid \mathcal{D})$$

Bayesian Model Comparison

Posterior Odds – How well M_1 estimates the underlying distribution compared to M_2

Prior Odds – How much prior beliefs favor M_1

Bayes Factor – How well data is predicted, i.e. marginal likelihood
Computing marginal likelihood using integration is sometimes intractable, can use stochastic approximations like:

- Monte Carlo
- Bayesian Monte Carlo
- Numerical Integration

Jeffreys-Lindley Paradox – Bayes factor favors the simpler model

$$\underbrace{\frac{p(M_1 \mid \mathcal{D})}{p(M_2 \mid \mathcal{D})}}_{\text{posterior odds}} = \frac{\frac{p(\mathcal{D} \mid M_1)p(M_1)}{p(\mathcal{D})}}{\frac{p(\mathcal{D} \mid M_2)p(M_2)}{p(\mathcal{D})}} = \underbrace{\frac{p(M_1)}{p(M_2)}}_{\text{prior odds}} \underbrace{\frac{p(\mathcal{D} \mid M_1)}{p(\mathcal{D} \mid M_2)}}_{\text{Bayes factor}} .$$

Model Selection Metrics

Akaike Information Criterion (AIC) – penalizes number of model parameters M

$$\log p(\boldsymbol{x} \mid \boldsymbol{\theta}) - M$$

Bayesian Information Criterion (BIC) – penalizes number of model parameters and complexity more heavily, relates it to number of samples N

$$\log p(\boldsymbol{x}) = \log \int p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \approx \log p(\boldsymbol{x} \mid \boldsymbol{\theta}) - \frac{1}{2} M \log N$$