Probabilistic models - Bayesian Methods

Mengye Ren

NYU

Oct 17, 2023

Overview

Why probabilistic modeling?

- A unified framework that covers many models, e.g., linear regression, logistic regression
- Learning as **statistical inference**
- Principled ways to incorporate your belief on the data generating distribution (inductive biases)

Two ways of generating data

- Two ways to model how the data is generated:
  - **Conditional**: $p(y \mid x)$
  - **Generative**: $p(x, y)$
- How to estimate the parameters of our model? Maximum likelihood estimation.
- Compare and contrast conditional and generative models.

Conditional models

Linear regression

Linear regression is one of the most important methods in machine learning and statistics.

Goal: Predict a real-valued **target** y (also called response) from a vector of **features** x (also called covariates).

Examples:

- Predicting house price given location, condition, build year etc.
- Predicting medical cost of a person given age, sex, region, BMI etc.
- Predicting age of a person based on their photos.

Problem setup

Data Training examples $\mathcal{D} = \{(x^{(n)}, y^{(n)})\}_{n=1}^{N}$, where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$.

Model A *linear* function h (parametrized by θ) to predict y from x:

$$h(x) = \sum_{i=0}^{d} \theta_i x_i = \theta^T x, \tag{1}$$

where $\theta \in \mathbb{R}^d$ are the **parameters** (also called weights).

Note that

- We incorporate the **bias term** (also called the intercept term) into x (i.e. $x_0 = 1$).
- We use superscript to denote the example id and subscript to denote the dimension id.

Parameter estimation

Loss function We estimate θ by minimizing the **squared loss** (the least square method):

$$J(\theta) = \frac{1}{N}\sum_{n=1}^{N}\left(y^{(n)} - \theta^T x^{(n)}\right)^2. \quad \text{(empirical risk)} \tag{2}$$

Matrix form

- Let $X \in \mathbb{R}^{N\times d}$ be the **design matrix** whose rows are input features.
- Let $\mathrm{y} \in \mathbb{R}^N$ be the vector of all targets.
- We want to solve

$$\hat{\theta} = \arg\min_{\theta}(X\theta - \mathrm{y})^T(X\theta - \mathrm{y}). \tag{3}$$

Solution Closed-form solution: $\hat{\theta} = (X^T X)^{-1} X^T \mathrm{y}$.

Review questions

- Derive the solution for linear regression.
- What if $X^T X$ is not invertible?

Review

We've seen

- Linear regression: response is a linear function of the inputs
- Estimate parameters by minimize the squared loss

But...

- Why squared loss is a reasonable choice for regression problems?
- What assumptions are we making on the data? (inductive bias)

Next,

- Derive linear regression from a probabilistic modeling perspective.

Assumptions in linear regression

- x and y are related through a linear function:

$$y = \theta^T x + \epsilon, \tag{4}$$

where ϵ is the **residual error** capturing all unmodeled effects (e.g., noise).

- The errors are distributed *iid* (independently and identically distributed):

$$\epsilon \sim \mathcal{N}(0, \sigma^2). \tag{5}$$

What's the distribution of $Y \mid X = x$?

$$p(y \mid x; \theta) = \mathcal{N}(\theta^T x, \sigma^2). \tag{6}$$

Imagine putting a Gaussian bump around the output of the linear predictor.

Maximum likelihood estimation (MLE)

Given a probabilistic model and a dataset $\mathcal{D}$, how to estimate the model parameters θ?

The **maximum likelihood principle** says that we should maximize the (conditional) likelihood of the data:

$$
\begin{aligned}
L(\theta) &\stackrel{\text{def}}{=} p(\mathcal{D}; \theta) && (7) \\
&= \prod_{n=1}^{N} p(y^{(n)} \mid x^{(n)}; \theta). && \text{(examples are distributed } iid\text{)} \quad (8)
\end{aligned}
$$

In practice, we maximize the **log likelihood** $\ell(\theta)$, or equivalently, minimize the negative log likelihood (NLL).

MLE for linear regression

Let's find the MLE solution for our model. Recall that $Y \mid X = x \sim \mathcal{N}(\theta^T x, \sigma^2)$.

$$
\begin{aligned}
\ell(\theta) &\stackrel{\text{def}}{=} \log L(\theta) && (9)\\
&= \log \prod_{n=1}^{N} p(y^{(n)} \mid x^{(n)}; \theta) && (10)\\
&= \sum_{n=1}^{N} \log p(y^{(n)} \mid x^{(n)}; \theta) && (11)\\
&= \sum_{n=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(n)} - \theta^T x^{(n)}\right)^2}{2\sigma^2} \right) && (12)\\
&= N \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{n=1}^{N} \left(y^{(n)} - \theta^T x^{(n)}\right)^2 && (13)
\end{aligned}
$$

Gradient of the likelihood

Recall that we obtained the normal equation by setting the derivative of the squared loss to zero. Now let's compute the derivative of the likelihood w.r.t. the parameters.

$$\ell(\theta) = N \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{n=1}^{N} \left(y^{(n)} - \theta^T x^{(n)} \right)^2 \tag{14}$$

$$\frac{\partial \ell}{\partial \theta_i} = -\frac{1}{\sigma^2} \sum_{n=1}^{N} (y^{(n)} - \theta^T x^{(n)}) x_i^{(n)}. \tag{15}$$

Review

We've seen

- Linear regression assumes that $Y \mid X = x$ follows a Gaussian distribution
- MLE of linear regression is equivalent to the least square method

However,

- Sometimes Gaussian distribution is not a reasonable assumption, e.g., classification
- Can we use the same modeling approach for other prediction tasks?

Next,

- Derive logistic regression for classification.

Assumptions in logistic regression

Consider binary classification where $Y \in \{0,1\}$. What should be the distribution $Y \mid X = x$?
We model $p(y \mid x)$ as a Bernoulli distribution:

$$p(y \mid x) = h(x)^y (1 - h(x))^{1-y}. \tag{16}$$

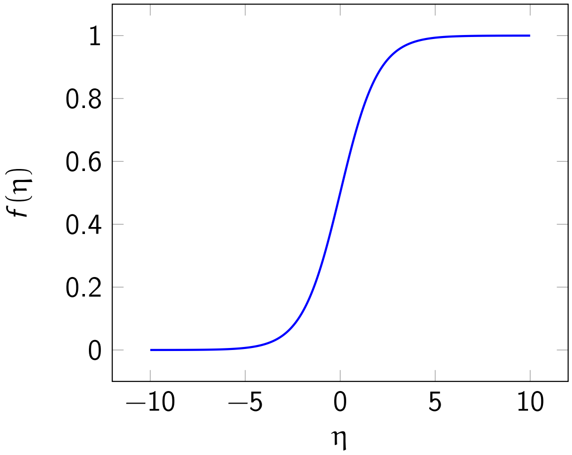
How should we parameterize $h(x)$?

- What is $p(y = 1 \mid x)$ and $p(y = 0 \mid x)$? $h(x) \in (0,1)$.
- What is the mean of $Y \mid X = x$? $h(x)$. (Think how we parameterize the mean in linear regression)
- Need a function f to map the linear predictor $\theta^T x$ in $\mathbb{R}$ to $(0,1)$:

$$f(\eta) = \frac{1}{1 + e^{-\eta}} \qquad \text{logistic function} \tag{17}$$

Logistic regression

$f(\eta) = \frac{1}{1+e^{-\eta}}$



- $p(y \mid x) = \text{Bernoulli}(f(\theta^T x))$.
- When do we have $p(y = 1 \mid x) = 1$ and $p(y = 0 \mid x) = 1$?
- Exercise: show that the **log odds** is

$$\log \frac{p(y = 1 \mid x)}{p(y = 0 \mid x)} = \theta^T x. \tag{18}$$

$$\implies \text{linear decision boundary} \tag{19}$$

- How do we extend it to multiclass classification? (more on this later)

MLE for logistic regression

Similar to linear regression, let's estimate θ by maximizing the conditional log likelihood.

$$\ell(\theta) = \sum_{n=1}^{N} \log p(y^{(n)} \mid x^{(n)}; \theta) \tag{20}$$

$$= \sum_{n=1}^{N} y^{(n)} \log f(\theta^T x^{(n)}) + (1 - y^{(n)}) \log(1 - f(\theta^T x^{(n)})) \tag{21}$$

- Closed-form solutions are not available.
- But, the likelihood is concave—gradient ascent gives us the unique optimal solution.

$$\theta := \theta + \alpha \nabla_\theta \ell(\theta). \tag{22}$$

Gradient descent for logistic regression

Math review: Chain rule

If z depends on y which itself depends on x, e.g., $z = (y(x))^2$, then $\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}$.

Likelihood for a single example: $\ell^n = y^{(n)} \log f(\theta^T x^{(n)}) + (1 - y^{(n)}) \log(1 - f(\theta^T x^{(n)}))$.

$$\frac{\partial \ell^n}{\partial \theta_i} = \frac{\partial \ell^n}{\partial f^n}\frac{\partial f^n}{\partial \theta_i} \tag{23}$$

$$= \left(\frac{y^{(n)}}{f^n} - \frac{1-y^{(n)}}{1-f^n}\right)\frac{\partial f^n}{\partial \theta_i} \qquad \frac{d}{dx}\ln x = \frac{1}{x} \tag{24}$$

$$= \left(\frac{y^{(n)}}{f^n} - \frac{1-y^{(n)}}{1-f^n}\right)\left(f^n(1-f^n)x_i^{(n)}\right) \qquad \text{Exercise: apply chain rule to } \frac{\partial f^n}{\partial \theta_i} \tag{25}$$

$$= (y^{(n)} - f^n)x_i^{(n)} \qquad \text{simplify by algebra} \tag{26}$$

The full gradient is thus $\frac{\partial \ell}{\partial \theta_i} = \sum_{n=1}^{N} (y^{(n)} - f(\theta^T x^{(n)}))x_i^{(n)}$.

A closer look at the gradient

$$\frac{\partial \ell}{\partial \theta_i} = \sum_{n=1}^{N} (y^{(n)} - f(\theta^T x^{(n)}))x_i^{(n)} \tag{27}$$

- Does this look familiar?
- Our derivation for linear regression and logistic regression are quite similar...
- Next, a more general family of models.

Compare linear regression and logistic regression

	linear regression	logistic regression
Combine the inputs	$\theta^T x$ (linear)	$\theta^T x$ (linear)
Output	real	categorical
Conditional distribution	Gaussian	Bernoulli
Transfer function $f(\theta^T x)$	identity	logistic
Mean $\mathbb{E}(Y \mid X = x; \theta)$	$f(\theta^T x)$	$f(\theta^T x)$

- x enters through a linear function.
- The main difference between the formulations is due to different conditional distributions.
- Can we generalize the idea to handle other output types, e.g., positive integers?

Construct a generalized regression model

Task: Given x, predict $p(y \mid x)$
Modeling:

- Choose a parametric family of distributions $p(y;\theta)$ with parameters $\theta \in \Theta$
- Choose a transfer function that maps a linear predictor in $\mathbb{R}$ to Θ

$$\underbrace{x}_{\in \mathrm{R}^d} \mapsto \underbrace{w^T x}_{\in \mathrm{R}} \mapsto \underbrace{f(w^T x)}_{\in \Theta} = \theta, \tag{28}$$

Learning: MLE: $\hat{\theta} \in \arg\max_\theta \log p(\mathcal{D};\hat{\theta})$
Inference: For prediction, use $x \to f(w^T x)$

Example: Construct Poisson regression

Say we want to predict the number of people entering a restaurant in New York during lunch time.

- What features would be useful?
- What's a good model for number of visitors (the output distribution)?

Math review: Poisson distribution

Given a random variable $Y \in 0, 1, 2, \ldots$ following Poisson(λ), we have

$$p(Y = k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \tag{29}$$

where $\lambda > 0$ and $\mathbb{E}[Y] = \lambda$.

The Poisson distribution is usually used to model the number of events occurring during a fixed period of time.

Example: Construct Poisson regression

We've decided that $Y \mid X = x \sim \text{Poisson}(\eta)$, what should be the transfer function f?

x enters linearly:

$$x \mapsto \underbrace{w^T x}_{\mathbf{R}} \mapsto \lambda = \underbrace{f(w^T x)}_{(0,\infty)}$$

Standard approach is to take

$$f(w^T x) = \exp\left(w^T x\right).$$

Likelihood of the full dataset $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$:

$$\log p(y_i; \lambda_i) = [y_i \log \lambda_i - \lambda_i - \log(y_i!)] \tag{30}$$

$$\log p(\mathcal{D}; w) = \sum_{i=1}^{n} \left[y_i \log\left[\exp\left(w^T x_i\right)\right] - \exp\left(w^T x_i\right) - \log(y_i!)\right] \tag{31}$$

$$= \sum_{i=1}^{n} \left[y_i w^T x_i - \exp\left(w^T x_i\right) - \log(y_i!)\right] \tag{32}$$

Multinomial Logistic Regression

- Say we want to get the predicted categorical distribution for a given $x \in \mathbb{R}^d$.
- First compute the scores ($\in \mathbb{R}^k$) and then their softmax:

$$x \mapsto \left(\langle w_1, x\rangle, \ldots, \langle w_k, x\rangle\right) \mapsto \theta = \left(\frac{\exp\left(w_1^T x\right)}{\sum_{i=1}^k \exp\left(w_i^T x\right)}, \ldots, \frac{\exp\left(w_k^T x\right)}{\sum_{i=1}^k \exp\left(w_i^T x\right)}\right)$$

- We can write the conditional probability for any $y \in \{1, \ldots, k\}$ as

$$p(y \mid x; w) = \frac{\exp\left(w_y^T x\right)}{\sum_{i=1}^k \exp\left(w_i^T x\right)}.$$

Review

Recipe for contructing a conditional distribution for prediction:

1. Define input and output space (as for any other model).
2. Choose the output distribution $p(y \mid x; \theta)$ based on the task
3. Choose the transfer function that maps $w^T x$ to a Θ.
4. (The formal family is called "generalized linear models".)

Learning:

- Fit the model by maximum likelihood estimation.
- Closed solutions do not exist in general, so we use gradient ascent.

Generative models

Review

We've seen

- Model the conditional distribution $p(y \mid x; \theta)$ using generalized linear models.
- (Previously) Directly map x to y, e.g., perceptron.

Next,

- Model the joint distribution $p(x, y; \theta)$.
- Predict the label for x as $\arg\max_{y \in \mathcal{Y}} p(x, y; \theta)$.

Generative modeling through the Bayes rule

Training:

$$p(x, y) = {\color{blue}p(x \mid y)}{\color{red}p(y)} \tag{33}$$

Testing:

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)} \qquad \text{Bayes rule} \tag{34}$$

$$\arg\max_y p(y \mid x) = \arg\max_y p(x \mid y)p(y) \tag{35}$$

Naive Bayes (NB) models

Let's consider binary text classification (e.g., fake vs genuine review) as a motivating example.

Bag-of-words representation of a document

- ["machine", "learning", "is", "fun", "."]
- $x_i \in \{0, 1\}$: whether the i-th word in our vocabulary exists in the input

$$x = [x_1, x_2, \ldots, x_d] \quad \text{where } d = \text{vocabulary size} \tag{36}$$

What's the probability of a document x?

$$
\begin{aligned}
p(x \mid y) &= p(x_1, \ldots, x_d \mid y) && (37)\\
&= p(x_1 \mid y)p(x_2 \mid y, x_1)p(x_3 \mid y, x_2, x_1) \ldots p(x_d \mid y, x_{d-1}, \ldots, x_1) \qquad \text{chain rule} && (38)\\
&= \prod_{i=1}^{d} p(x_i \mid y, x_{<i}) && (39)
\end{aligned}
$$

Naive Bayes assumption

Challenge: $p(x_i \mid y, x_{<i})$ is hard to model (and estimate), especially for large i.
Solution:

Naive Bayes assumption

Features are **conditionally independent** given the label:

$$p(x \mid y) = \prod_{i=1}^{d} p(x_i \mid y). \tag{40}$$

A strong assumption in general, but works well in practice.

Parametrize $p(x_i \mid y)$ and $p(y)$

For binary x_i, assume $p(x_i \mid y)$ follows Bernoulli distributions.

$$p(x_i = 1 \mid y = 1) = \theta_{i,1}, \;\; p(x_i = 1 \mid y = 0) = \theta_{i,0}. \tag{41}$$

Similarly,

$$p(y = 1) = \theta_0. \tag{42}$$

Thus,

$$\begin{aligned}
p(x, y) &= p(x \mid y) p(y) && (43)\\
&= p(y) \prod_{i=1}^{d} p(x_i \mid y) && \text{NB assumption} \quad (44)\\
&= p(y) \prod_{i=1}^{d} \theta_{i,y} \mathbb{I}\{x_i = 1\} + (1 - \theta_{i,y}) \mathbb{I}\{x_i = 0\} && (45)
\end{aligned}$$

Indicator function $\mathbb{I}\{\text{condition}\}$ evaluates to 1 if "condition" is true and 0 otherwise.

MLE for our NB model

We maximize the likelihood of the data $\prod_{n=1}^{N} p_\theta(x^{(n)}, y^{(n)})$ (as opposed to the *conditional* likelihood we've seen before).

$$\frac{\partial}{\partial\theta_{j,1}}\ell = \frac{\partial}{\partial\theta_{j,1}}\sum_{n=1}^{N}\sum_{i=1}^{d}\log\left(\theta_{i,y^{(n)}}\mathbb{I}\left\{x_i^{(n)}=1\right\}+\left(1-\theta_{i,y^{(n)}}\right)\mathbb{I}\left\{x_i^{(n)}=0\right\}\right)+\log p_{\theta_0}(y^{(n)}) \tag{46}$$

$$= \frac{\partial}{\partial\theta_{j,1}}\sum_{n=1}^{N}\log\left(\theta_{j,y^{(n)}}\mathbb{I}\left\{x_j^{(n)}=1\right\}+\left(1-\theta_{j,y^{(n)}}\right)\mathbb{I}\left\{x_j^{(n)}=0\right\}\right) \qquad \text{ignore } i \neq j \tag{47}$$

$$= \sum_{n=1}^{N}\mathbb{I}\left\{y^{(n)}=1\wedge x_j^{(n)}=1\right\}\frac{1}{\theta_{j,1}}+\mathbb{I}\left\{y^{(n)}=1\wedge x_j^{(n)}=0\right\}\frac{1}{1-\theta_{j,1}} \qquad \text{ignore } y^{(n)}=0 \tag{48}$$

MLE solution for our NB model

Set $\frac{\partial}{\partial \theta_{j,1}} \ell$ to zero:

$$\theta_{j,1} = \frac{\sum_{n=1}^{N} \mathbb{I}\left\{y^{(n)} = 1 \wedge x_j^{(n)} = 1\right\}}{\sum_{n=1}^{N} \mathbb{I}\left\{y^{(n)} = 1\right\}} \tag{49}$$

In practice, count words:

$$\frac{\text{number of fake reviews containing ``absolutely''}}{\text{number of fake reviews}}$$

Exercise: show that

$$\theta_{j,0} = \frac{\sum_{n=1}^{N} \mathbb{I}\left\{y^{(n)} = 0 \wedge x_j^{(n)} = 1\right\}}{\sum_{n=1}^{N} \mathbb{I}\left\{y^{(n)} = 0\right\}} \tag{50}$$

$$\theta_0 = \frac{\sum_{n=1}^{N} \mathbb{I}\left\{y^{(n)} = 1\right\}}{N} \tag{51}$$

Review

NB assumption: conditionally independent features given the label

Recipe for learning a NB model:

1. Choose $p(x_i \mid y)$, e.g., Bernoulli distribution for binary x_i.
2. Choose $p(y)$, often a categorical distribution.
3. Estimate parameters by MLE (same as the strategy for conditional models) .

Next, NB with continuous features.

NB with continuous inputs

Let's consider a multiclass classification task with continuous inputs.

$$p(x_i \mid y) \sim \mathcal{N}(\mu_{i,y}, \sigma^2_{i,y}) \tag{52}$$

$$p(y = k) = \theta_k \tag{53}$$

Likelihood of the data:

$$p(\mathcal{D}) = \prod_{n=1}^{N} p(y^{(n)}) \prod_{i=1}^{d} p(x_i^{(n)} \mid y^{(n)}) \tag{54}$$

$$= \prod_{n=1}^{N} \theta_{y^{(n)}} \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_{i,y^{(n)}}} \exp\left(-\frac{1}{2\sigma^2_{i,y^{(n)}}}\left(x_i^{(n)} - \mu_{i,y^{(n)}}\right)^2\right) \tag{55}$$

MLE for Gaussian NB

Log likelihood:

$$\ell = \sum_{n=1}^{N} \log \theta_{y^{(n)}} + \sum_{n=1}^{N} \sum_{i=1}^{d} \log \frac{1}{\sqrt{2\pi}\sigma_{i,y^{(n)}}} - \frac{1}{2\sigma_{i,y^{(n)}}^2} \left(x_i^{(n)} - \mu_{i,y^{(n)}} \right)^2 \tag{56}$$

$$\frac{\partial}{\partial \mu_{j,k}} \ell = \frac{\partial}{\partial \mu_{j,k}} \sum_{n: y^{(n)}=k} -\frac{1}{2\sigma_{j,k}^2} \left(x_j^{(n)} - \mu_{j,k} \right)^2 \qquad \text{ignore irrelevant terms} \tag{57}$$

$$= \sum_{n: y^{(n)}=k} \frac{1}{\sigma_{j,k}^2} \left(x_j^{(n)} - \mu_{j,k} \right) \tag{58}$$

Set $\frac{\partial}{\partial \mu_{j,k}} \ell$ to zero:

$$\mu_{j,k} = \frac{\sum_{n: y^{(n)}=k} x_j^{(n)}}{\sum_{n: y^{(n)}=k} 1} = \text{sample mean of } x_j \text{ in class } k \tag{59}$$

MLE for Gaussian NB

Show that

$$\sigma^2_{j,k} = \frac{\sum_{n:y^{(n)}=k}\left(x_j^{(n)} - \mu_{j,k}\right)^2}{\sum_{n:y^{(n)}=k} 1} = \text{sample variance of } x_j \text{ in class } k \tag{60}$$

$$\theta_k = \frac{\sum_{n:y^{(n)}=k} 1}{N} \qquad \text{(class prior)} \tag{61}$$

Decision boundary of the Gaussian NB model

Is the Gaussian NB model a linear classifier?

$$\log \frac{p(y=1 \mid x)}{p(y=0 \mid x)} = \log \frac{p(x \mid y=1)p(y=1)}{p(x \mid y=0)p(y=0)} \tag{62}$$

$$= \log \frac{\theta_0}{1-\theta_0} + \sum_{i=1}^{d} \left(\log \sqrt{\frac{\sigma_{i,0}^2}{\sigma_{i,1}^2}} + \left(\frac{(x_i - \mu_{i,0})^2}{2\sigma_{i,0}^2} - \frac{(x_i - \mu_{i,1})^2}{2\sigma_{i,1}^2} \right) \right) \quad \text{quadratic} \tag{63}$$

$$\text{assume that } \sigma_{i,0} = \sigma_{i,1} = \sigma_i, \; (\theta_0 = 0.5) \tag{64}$$

$$= \sum_{i=1}^{d} \frac{1}{2\sigma_i^2} \left((x_i - \mu_{i,0})^2 - (x_i - \mu_{i,1})^2 \right) \tag{65}$$

$$= \sum_{i=1}^{d} \frac{\mu_{i,1} - \mu_{i,0}}{\sigma_i^2} x_i + \frac{\mu_{i,0}^2 - \mu_{i,1}^2}{2\sigma_i^2} \quad \text{linear} \tag{66}$$

Decision boundary of the Gaussian NB model

Assuming the variance of each feature is the same for both classes, we have

$$
\begin{aligned}
\log \frac{p(y=1 \mid x)}{p(y=0 \mid x)} &= \sum_{i=1}^{d} \frac{\mu_{i,1}-\mu_{i,0}}{\sigma_i^2} x_i + \frac{\mu_{i,0}^2-\mu_{i,1}^2}{2\sigma_i^2} && (67) \\
&= \theta^T x \qquad \text{where else have we seen it?} && (68) \\
& && (69)
\end{aligned}
$$

$$
\begin{aligned}
\theta_i &= \frac{\mu_{i,1}-\mu_{i,0}}{\sigma_i^2} \qquad \text{for } i \in [1, d] && (70) \\
\theta_0 &= \sum_{i=1}^{d} \frac{\mu_{i,0}^2-\mu_{i,1}^2}{2\sigma_i^2} \qquad \text{bias term} && (71)
\end{aligned}
$$

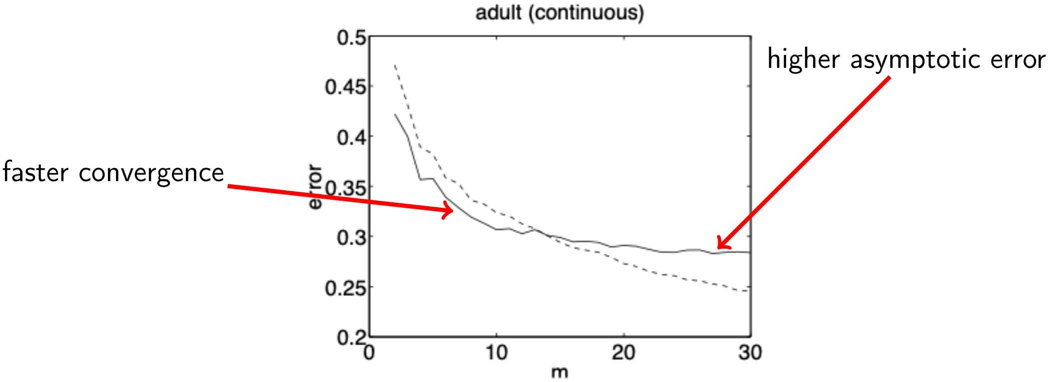
Naive Bayes vs logistic regression

	logistic regression	Gaussian naive Bayes
model type	conditional/discriminative	generative
parametrization	$p(y \mid x)$	$p(x \mid y)$, $p(y)$
assumptions on Y	Bernoulli	Bernoulli
assumptions on X	—	Gaussian
decision boundary	$\theta_{\text{LR}}^T x$	$\theta_{\text{GNB}}^T x$

Given the same training data, is $\theta_{\text{LR}} = \theta_{\text{GNB}}$?

Generative vs discriminative classifiers

Ng, A. and Jordan, M. (2002). On discriminative versus generative classifiers: A comparison of logistic regression and naive Bayes. In Advances in Neural Information Processing Systems 14.



Solid line: naive Bayes; dashed line: logistic regression.

Naive Bayes vs logistic regression

Logistic regression and Gaussian naive Bayes converge to the same classifier asymptotically, assuming the GNB assumption holds.

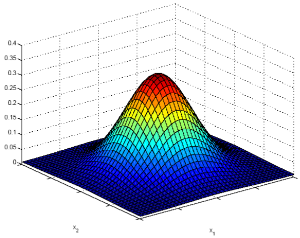
- Data points are generated from Gaussian distributions for each class
- Each dimension is independently generated
- Shared variance

What if the GNB assumption is not true?

Multivariate Gaussian Distribution

- $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, a Gaussian (or normal) distribution defined as

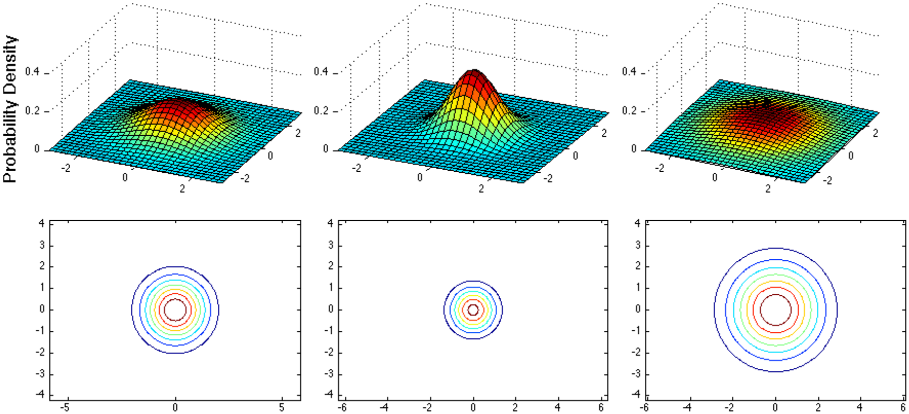
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$



- Mahalanobis distance $(\mathbf{x}-\boldsymbol{\mu}_k)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)$ measures the distance from $\mathbf{x}$ to $\boldsymbol{\mu}$ in terms of Σ
- It normalizes for difference in variances and correlations

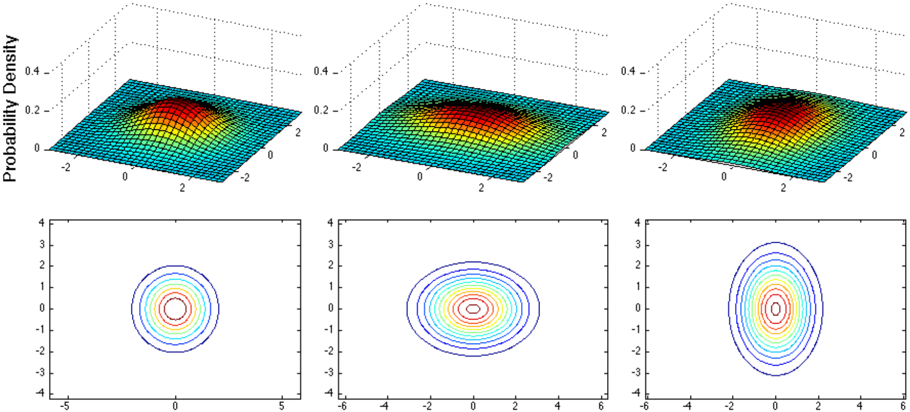
Bivariate Normal

$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ $\Sigma = 0.5 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ $\Sigma = 2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

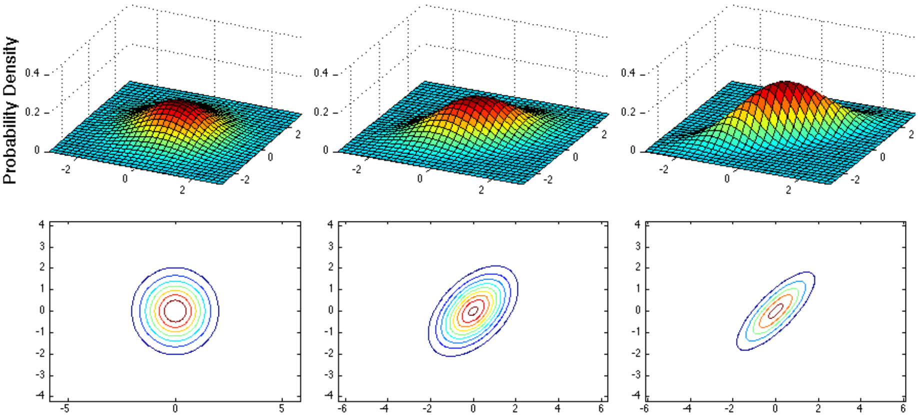


Bivariate Normal

$var(x_1) = var(x_2)$ $\quad$ $var(x_1) > var(x_2)$ $\quad$ $var(x_1) < var(x_2)$

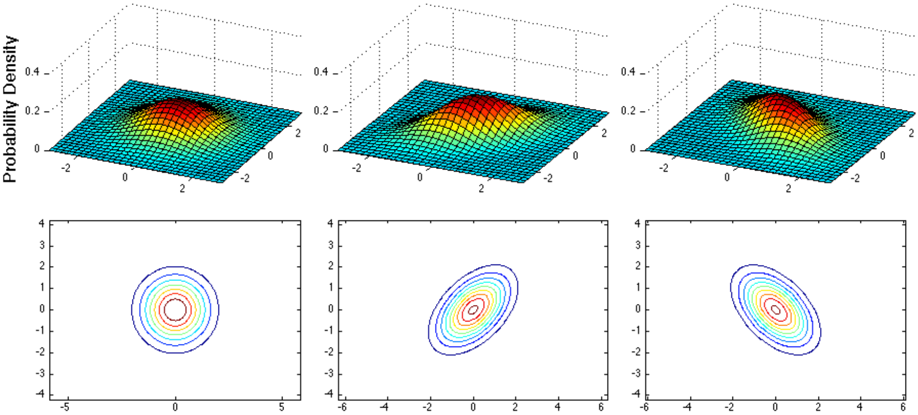


$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$



Bivariate Normal

$Cov(x_1, x_2) = 0$ $\quad$ $Cov(x_1, x_2) > 0$ $\quad$ $Cov(x_1, x_2) < 0$



Gaussian Bayes Classifier

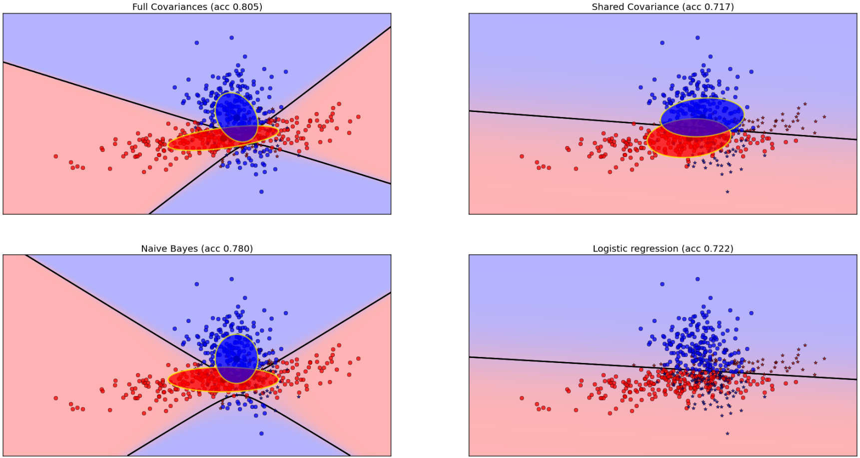
- Gaussian Bayes Classifier in its general form assumes that $p(\mathbf{x}|y)$ is distributed according to a multivariate normal (Gaussian) distribution
- Multivariate Gaussian distribution:

$$p(\mathbf{x}|t=k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \Sigma_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right]$$

where $|\Sigma_k|$ denotes the determinant of the matrix, and d is dimension of $\mathbf{x}$

- Each class k has associated mean vector $\boldsymbol{\mu}_k$ and covariance matrix Σ_k
- Σ_k has $\mathcal{O}(d^2)$ parameters - could be hard to estimate

Example


Full Covariances (acc 0.805)
Shared Covariance (acc 0.717)
Naive Bayes (acc 0.780)
Logistic regression (acc 0.722)

Gaussian Bayes Binary Classifier Cases

- Full covariance: Quadratic decision boundary
- Shared covariance: Linear decision boundary
- Naive Bayes: Diagonal covariance matrix, quadratic decision boundary
- If data is truly Gaussian distributed, then shared covariance = logistic regression.
- But logistic regression can learn other distributions.

Bayesian ML: Classical Statistics

Parametric Family of Densities

- A **parametric family of densities** is a set

$$\{p(y \mid \theta) : \theta \in \Theta\},$$

  - where $p(y \mid \theta)$ is a density on a **sample space** $\mathcal{Y}$, and
  - θ is a **parameter** in a [finite dimensional] **parameter space** Θ.
- This is the common starting point for a treatment of classical or Bayesian statistics.
- In this lecture, whenever we say "density", we could replace it with "mass function." (and replace integrals with sums).

Frequentist or "Classical" Statistics

- We're still working with a parametric family of densities:

$$\{p(y \mid \theta) \mid \theta \in \Theta\}.$$

- Assume that $p(y \mid \theta)$ governs the world we are observing, for some $\theta \in \Theta$.
- If we knew the right $\theta \in \Theta$, there would be no need for statistics.
- But instead of θ, we have data $\mathcal{D}$: $y_1, \ldots, y_n$ sampled i.i.d. from $p(y \mid \theta)$.
- Statistics is about how to get by with $\mathcal{D}$ in place of θ.

Point Estimation

- One type of statistical problem is **point estimation**.
- A **statistic** $s = s(\mathcal{D})$ is any function of the data.
- A statistic $\hat{\theta} = \hat{\theta}(\mathcal{D})$ taking values in Θ is a **point estimator of** θ.
- A good point estimator will have $\hat{\theta} \approx \theta$.
- **Desirable statistical properties of point estimators**:
    - **Consistency:** As data size $n \to \infty$, we get $\hat{\theta}_n \to \theta$.
    - **Efficiency:** (Roughly speaking) $\hat{\theta}_n$ is as accurate as we can get from a sample of size n.
- **Maximum likelihood estimators** are consistent and efficient under reasonable conditions.

Example of Point Estimation: Coin Flipping

- Parametric family of mass functions:

$$p(\text{Heads} \mid \theta) = \theta,$$

for $\theta \in \Theta = (0, 1)$.

Coin Flipping: MLE

- Data $\mathcal{D} = (H, H, T, T, T, T, T, H, \ldots, T)$, assumed i.i.d. flips.
  - n_h: number of heads
  - n_t: number of tails
- **Likelihood function** for data $\mathcal{D}$:

$$L_{\mathcal{D}}(\theta) = p(\mathcal{D} \mid \theta) = \theta^{n_h}(1-\theta)^{n_t}$$

- As usual, it is easier to maximize the log-likelihood function:

$$\begin{aligned}\hat{\theta}_{\text{MLE}} &= \underset{\theta \in \Theta}{\arg\max} \log L_{\mathcal{D}}(\theta) \\ &= \underset{\theta \in \Theta}{\arg\max} \left[n_h \log \theta + n_t \log(1-\theta)\right]\end{aligned}$$

- First order condition (equating the derivative to zero):

$$\frac{n_h}{\theta} - \frac{n_t}{1-\theta} = 0 \iff \theta = \frac{n_h}{n_h + n_t}$$

$\hat{\theta}_{\text{MLE}}$ is the empirical fraction of heads.

Bayesian Statistics: Introduction

Bayesian Statistics

- Baysian statistics introduces a crucial new ingredient: the **prior distribution.**
- A **prior distribution** $p(\theta)$ is a distribution on the parameter space Θ.
- The prior reflects our belief about θ, **before seeing any data**.

A Bayesian Model

- A [parametric] Bayesian model consists of two pieces:
  1. A parametric family of densities
$$\{p(\mathcal{D} \mid \theta) \mid \theta \in \Theta\}.$$
  2. A **prior distribution** $p(\theta)$ on parameter space Θ.
- Putting the pieces together, we get a joint density on θ and $\mathcal{D}$:
$$p(\mathcal{D}, \theta) = p(\mathcal{D} \mid \theta) p(\theta).$$

The Posterior Distribution

- The **posterior distribution** for θ is $p(\theta \mid \mathcal{D})$.
- Whereas the prior represents belief about θ before observing data $\mathcal{D}$,
- The posterior represents the **rationally updated belief** about θ, after seeing $\mathcal{D}$.

Expressing the Posterior Distribution

- By Bayes rule, can write the posterior distribution as

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

- Let's consider both sides as functions of θ, for fixed $\mathcal{D}$.
- Then both sides are densities on Θ and we can write

$$\underbrace{p(\theta \mid \mathcal{D})}_{\text{posterior}} \propto \underbrace{p(\mathcal{D} \mid \theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}.$$

- Where $\propto$ means we've dropped factors that are independent of θ.

Coin Flipping: Bayesian Model

- Recall that we have a parametric family of mass functions:

$$p(\text{Heads} \mid \theta) = \theta,$$

  for $\theta \in \Theta = (0,1)$.

- We need a prior distribution $p(\theta)$ on $\Theta = (0,1)$.
- One convenient choice would be a distribution from the Beta family

- **Prior:**

$$
\begin{aligned}
\theta &\sim \mathrm{Beta}(\alpha, \beta) \\
p(\theta) &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}
\end{aligned}
$$

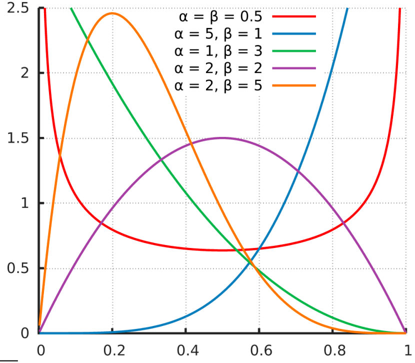


Figure by Horas based on the work of Krishnavedala (Own work) [Public domain], via Wikimedia Commons http://commons.wikimedia.org/wiki/File:Beta_distribution_pdf.svg.

Coin Flipping: Beta Prior

- **Prior:**

$$
\begin{aligned}
\theta &\sim \text{Beta}(h,t) \\
p(\theta) &\propto \theta^{h-1}(1-\theta)^{t-1}
\end{aligned}
$$

- **Mean of Beta distribution:**

$$
\mathbb{E}\theta = \frac{h}{h+t}
$$

- **Mode of Beta distribution:**

$$
\arg\max_{\theta} p(\theta) = \frac{h-1}{h+t-2}
$$

for $h,t > 1$.

Coin Flipping: Posterior

- **Prior:**

$$\begin{aligned}\theta &\sim \text{Beta}(h,t)\\ p(\theta) &\propto \theta^{h-1}(1-\theta)^{t-1}\end{aligned}$$

- **Likelihood function**

$$L(\theta) = p(\mathcal{D} \mid \theta) = \theta^{n_h}(1-\theta)^{n_t}$$

- **Posterior density:**

$$\begin{aligned}p(\theta \mid \mathcal{D}) &\propto p(\theta)p(\mathcal{D} \mid \theta)\\ &\propto \theta^{h-1}(1-\theta)^{t-1} \times \theta^{n_h}(1-\theta)^{n_t}\\ &= \theta^{h-1+n_h}(1-\theta)^{t-1+n_t}\end{aligned}$$

The Posterior is in the Beta Family!

- **Prior:**

$$\begin{aligned}\theta &\sim \text{Beta}(h,t)\\ p(\theta) &\propto \theta^{h-1}(1-\theta)^{t-1}\end{aligned}$$

- **Posterior density:**

$$p(\theta \mid \mathcal{D}) \propto \theta^{h-1+n_h}(1-\theta)^{t-1+n_t}$$

- **Posterior is in the beta family**:

$$\theta \mid \mathcal{D} \sim \text{Beta}(h+n_h, t+n_t)$$

- **Interpretation**:
    - Prior initializes our counts with h heads and t tails.
    - Posterior increments counts by observed n_h and n_t.

- In this case, the posterior is in the same distribution family as the prior.
- Let π be a family of prior distributions on Θ.
- Let P parametric family of distributions with parameter space Θ.

Definition

A family of distributions π **is conjugate to** parametric model P if for any prior in π, the posterior is always in π.

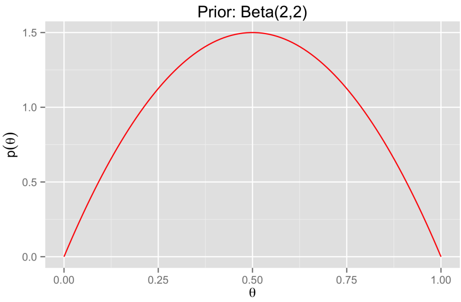
- The beta family is conjugate to the coin-flipping (i.e. Bernoulli) model.

Coin Flipping: Concrete Example

- Suppose we have a coin, possibly biased (**parametric probability model**):

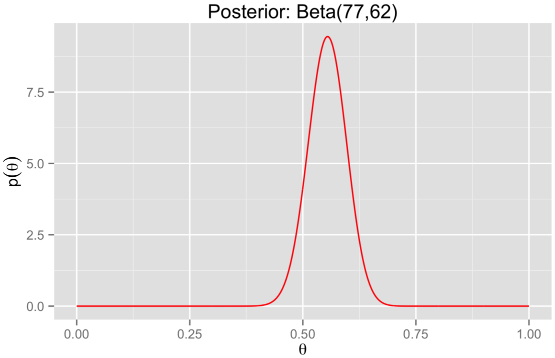
$$p(\text{Heads} \mid \theta) = \theta.$$

- **Parameter space** $\theta \in \Theta = [0,1]$.
- **Prior distribution:** $\theta \sim \text{Beta}(2,2)$.



Example: Coin Flipping

- Next, we gather some data $\mathcal{D} = \{H, H, T, T, T, T, T, H, \ldots, T\}$:
- Heads: 75   Tails: 60
  - $\hat{\theta}_{\text{MLE}} = \frac{75}{75+60} \approx 0.556$
- **Posterior distribution:** $\theta \mid \mathcal{D} \sim \textbf{Beta}(77, 62)$:



Bayesian Point Estimates

- We have the posterior distribution $\theta \mid \mathcal{D}$.
- What if someone asks us for a point estimate $\hat{\theta}$ for θ?
- Common options:
  - **posterior mean** $\hat{\theta} = \mathbb{E}[\theta \mid \mathcal{D}]$
  - **maximum a posteriori (MAP) estimate** $\hat{\theta} = \arg\max_\theta p(\theta \mid \mathcal{D})$
    - Note: this is the **mode** of the posterior distribution

What else can we do with a posterior?

- Look at it: display uncertainty estimates to our client
- Extract a **credible set** for θ (a Bayesian confidence interval).
    - e.g. Interval $[a, b]$ is a 95% **credible set** if

$$\mathbb{P}(\theta \in [a, b] \mid \mathcal{D}) \geqslant 0.95$$

- Select a point estimate using **Bayesian decision theory**:
    - Choose a loss function.
    - Find action **minimizing expected risk w.r.t. posterior**

Bayesian Decision Theory

Bayesian Decision Theory

- Ingredients:
  - **Parameter space** Θ.
  - **Prior**: Distribution $p(\theta)$ on Θ.
  - **Action space** $\mathcal{A}$.
  - **Loss function**: $\ell : \mathcal{A} \times \Theta \to \mathrm{R}$.
- The **posterior risk** of an action $a \in \mathcal{A}$ is

$$\begin{aligned} r(a) &:= \mathbb{E}\left[\ell(\theta, a) \mid \mathcal{D}\right] \\ &= \int \ell(\theta, a) p(\theta \mid \mathcal{D})\, d\theta. \end{aligned}$$

  - It's the **expected loss under the posterior.**
- A **Bayes action** a^* is an action that minimizes posterior risk:

$$r(a^*) = \min_{a \in \mathcal{A}} r(a)$$

Bayesian Point Estimation

- General Setup:
  - Data $\mathcal{D}$ generated by $p(y \mid \theta)$, for unknown $\theta \in \Theta$.
  - We want to produce a **point estimate** for θ.
- Choose:
  - **Prior** $p(\theta)$ on $\Theta = \mathrm{R}$.
  - **Loss** $\ell(\hat{\theta}, \theta)$
- Find **action** $\hat{\theta} \in \Theta$ that minimizes the **posterior risk:**

$$\begin{aligned} r(\hat{\theta}) &= \mathbb{E}\left[\ell(\hat{\theta}, \theta) \mid \mathcal{D}\right] \\ &= \int \ell(\hat{\theta}, \theta) p(\theta \mid \mathcal{D})\, d\theta \end{aligned}$$

Important Cases

- Squared Loss : $\ell(\hat{\theta}, \theta) = \left(\theta - \hat{\theta}\right)^2 \quad \Rightarrow$ posterior mean
- Zero-one Loss: $\ell(\theta, \hat{\theta}) = \mathbb{1}[\theta \neq \hat{\theta}] \quad \Rightarrow$ posterior mode
- Absolute Loss : $\ell(\hat{\theta}, \theta) = \left|\theta - \hat{\theta}\right| \quad \Rightarrow$ posterior median

Bayesian Point Estimation: Square Loss

- Find **action** $\hat{\theta} \in \Theta$ that minimizes **posterior risk**

$$r(\hat{\theta}) = \int \left(\theta - \hat{\theta}\right)^2 p(\theta \mid \mathcal{D})\, d\theta.$$

- Differentiate:

$$\begin{aligned}
\frac{dr(\hat{\theta})}{d\hat{\theta}} &= -\int 2\left(\theta - \hat{\theta}\right) p(\theta \mid \mathcal{D})\, d\theta \\
&= -2\int \theta p(\theta \mid \mathcal{D})\, d\theta + 2\hat{\theta} \underbrace{\int p(\theta \mid \mathcal{D})\, d\theta}_{=1} \\
&= -2\int \theta p(\theta \mid \mathcal{D})\, d\theta + 2\hat{\theta}
\end{aligned}$$

Bayesian Point Estimation: Square Loss

- Derivative of posterior risk is

$$\frac{dr(\hat{\theta})}{d\hat{\theta}} = -2\int \theta p(\theta \mid \mathcal{D})\, d\theta + 2\hat{\theta}.$$

- First order condition $\frac{dr(\hat{\theta})}{d\hat{\theta}} = 0$ gives

$$\begin{aligned} \hat{\theta} &= \int \theta p(\theta \mid \mathcal{D})\, d\theta \\ &= \mathbb{E}[\theta \mid \mathcal{D}] \end{aligned}$$

- The **Bayes action** for **square loss** is the posterior mean.

Interim summary

Recap and Interpretation

- The prior represents belief about θ before observing data $\mathcal{D}$.
- The posterior represents **rationally updated beliefs** after seeing $\mathcal{D}$.
- All inferences and action-taking are based on the posterior distribution.
- In the Bayesian approach,
  - No issue of justifying an estimator.
  - Only choices are
    - **family of distributions**, indexed by Θ, and
    - **prior distribution** on Θ
  - For decision making, we need a **loss function**.

Recap: Conditional Probability Models

Conditional Probability Modeling

- **Input space** $\mathcal{X}$
- **Outcome space** $\mathcal{Y}$
- **Action space** $\mathcal{A} = \{p(y) \mid p \text{ is a probability distribution on } \mathcal{Y}\}$.
- **Hypothesis space** $\mathcal{F}$ contains prediction functions $f : \mathcal{X} \to \mathcal{A}$.
- **Prediction function** $f \in \mathcal{F}$ takes input $x \in \mathcal{X}$ and produces a **distribution** on $\mathcal{Y}$
- A **parametric family of conditional densities** is a set

$$\{p(y \mid x, \theta) : \theta \in \Theta\},$$

  - where $p(y \mid x, \theta)$ is a density on **outcome space** $\mathcal{Y}$ for each x in **input space** $\mathcal{X}$, and
  - θ is a **parameter** in a [finite dimensional] **parameter space** Θ.
- This is the common starting point for either classical or Bayesian regression.

Classical treatment: Likelihood Function

- **Data:** $\mathcal{D} = (y_1, \ldots, , y_n)$
- The probability density for our data $\mathcal{D}$ is

$$p(\mathcal{D} \mid x_1, \ldots, x_n, \theta) \quad = \quad \prod_{i=1}^{n} p(y_i \mid x_i, \theta).$$

- For fixed $\mathcal{D}$, the function $\theta \mapsto p(\mathcal{D} \mid x, \theta)$ is the **likelihood function**:

$$L_{\mathcal{D}}(\theta) = p(\mathcal{D} \mid x, \theta),$$

  where $x = (x_1, \ldots, x_n)$.

Maximum Likelihood Estimator

- The **maximum likelihood estimator (MLE)** for θ in the family $\{p(y \mid x, \theta) \mid \theta \in \Theta\}$ is

$$\hat{\theta}_{\text{MLE}} = \underset{\theta \in \Theta}{\arg\max}\, L_{\mathcal{D}}(\theta).$$

- MLE corresponds to ERM, if we set the loss to be the negative log-likelihood.
- The corresponding prediction function is

$$\hat{f}(x) = p(y \mid x, \hat{\theta}_{\text{MLE}}).$$

Bayesian Conditional Probability Models

Bayesian Conditional Models

- Input space $\mathcal{X} = \mathrm{R}^d$ Outcome space $\mathcal{Y} = \mathrm{R}$
- The Bayesian conditional model has two components:
  - A **parametric family of conditional densities**:

$$\{p(y \mid x, \theta) : \theta \in \Theta\}$$

  - A **prior distribution** $p(\theta)$ on $\theta \in \Theta$.

The Posterior Distribution

- The **prior distribution** $p(\theta)$ represents our beliefs about θ before seeing $\mathcal{D}$.
- The **posterior distribution** for θ is

$$
\begin{aligned}
p(\theta \mid \mathcal{D}, x) &\propto p(\mathcal{D} \mid \theta, x) p(\theta) \\
&= \underbrace{L_{\mathcal{D}}(\theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}
\end{aligned}
$$

- Posterior represents the **rationally updated beliefs** after seeing $\mathcal{D}$.
- Each θ corresponds to a prediction function,
  - i.e. the conditional distribution function $p(y \mid x, \theta)$.

Point Estimates of Parameter

- What if we want point estimates of θ?
- We can use **Bayesian decision theory** to derive point estimates.
- We may want to use
  - $\hat{\theta} = \mathbb{E}[\theta \mid \mathcal{D}, x]$ (the posterior mean estimate)
  - $\hat{\theta} = \text{median}[\theta \mid \mathcal{D}, x]$
  - $\hat{\theta} = \arg\max_{\theta \in \Theta} p(\theta \mid \mathcal{D}, x)$ (the MAP estimate)
- depending on our loss function.

Back to the basic question - Bayesian Prediction Function

- Find a function takes input $x \in \mathcal{X}$ and produces a **distribution** on $\mathcal{Y}$
- In the frequentist approach:
  - Choose family of conditional probability densities (hypothesis space).
  - Select one conditional probability from family, e.g. using MLE.
- In the Bayesian setting:
  - We choose a parametric family of conditional densities

$$\{p(y \mid x, \theta) : \theta \in \Theta\},$$

  - and a prior distribution $p(\theta)$ on this set.
- Having set our Bayesian model, how do we predict a distribution on y for input x?
- We don't need to make a discrete selection from the hypothesis space: we **maintain uncertainty**.

The Prior Predictive Distribution

- Suppose we have not yet observed any data.
- In the Bayesian setting, we can still produce a prediction function.
- The **prior predictive distribution** is given by

$$x \mapsto p(y \mid x) = \int p(y \mid x; \theta) p(\theta)\, d\theta.$$

- This is an average of all conditional densities in our family, weighted by the prior.

The Posterior Predictive Distribution

- Suppose we've already seen data $\mathcal{D}$.
- The **posterior predictive distribution** is given by

$$x \mapsto p(y \mid x, \mathcal{D}) = \int p(y \mid x; \theta) p(\theta \mid \mathcal{D})\, d\theta.$$

- This is an average of all conditional densities in our family, weighted by the posterior.

Comparison to Frequentist Approach

- In Bayesian statistics we have two distributions on Θ:
  - the prior distribution $p(\theta)$
  - the posterior distribution $p(\theta \mid \mathcal{D})$.
- These distributions over parameters correspond to distributions on the hypothesis space:

$$\{p(y \mid x, \theta) : \theta \in \Theta\}.$$

- In the frequentist approach, we choose $\hat{\theta} \in \Theta$, and predict

$$p(y \mid x, \hat{\theta}(\mathcal{D})).$$

- In the Bayesian approach, we integrate out over Θ w.r.t. $p(\theta \mid \mathcal{D})$ and predict with

$$p(y \mid x, \mathcal{D}) = \int p(y \mid x; \theta) p(\theta \mid \mathcal{D}) \, d\theta$$

What if we don't want a full distribution on y?

- Once we have a predictive distribution $p(y \mid x, \mathcal{D})$,
  - we can easily generate single point predictions.
- $x \mapsto \mathbb{E}[y \mid x, \mathcal{D}]$, to minimize expected square error.
- $x \mapsto \text{median}[y \mid x, \mathcal{D}]$, to minimize expected absolute error
- $x \mapsto \arg\max_{y \in \mathcal{Y}} p(y \mid x, \mathcal{D})$, to minimize expected 0/1 loss
- Each of these can be derived from $p(y \mid x, \mathcal{D})$.

Gaussian Regression Example

Example in 1-Dimension: Setup

- Input space $\mathcal{X} = [-1, 1]$ Output space $\mathcal{Y} = \mathbf{R}$
- Given x, the world generates y as

$$y = w_0 + w_1 x + \varepsilon,$$

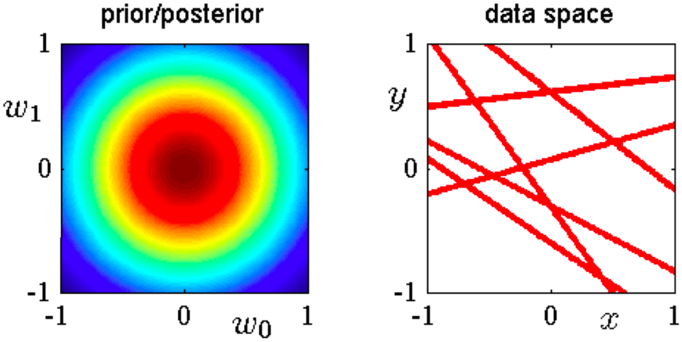
where $\varepsilon \sim \mathcal{N}(0, 0.2^2)$.

- Written another way, the **conditional probability model** is

$$y \mid x, w_0, w_1 \sim \mathcal{N}\left(w_0 + w_1 x\,,\, 0.2^2\right).$$

- What's the parameter space? $\mathbf{R}^2$.
- **Prior distribution:** $w = (w_0, w_1) \sim \mathcal{N}\left(0, \frac{1}{2} I\right)$

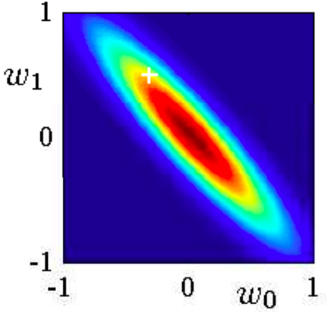
- **Prior distribution:** $w = (w_0, w_1) \sim \mathcal{N}\left(0, \frac{1}{2}I\right)$ (Illustrated on left)

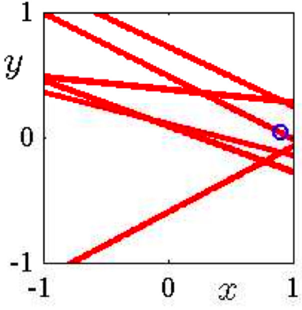


- On right, $y(x) = \mathbb{E}[y \mid x, w] = w_0 + w_1 x$, for randomly chosen $w \sim p(w) = \mathcal{N}\left(0, \frac{1}{2}I\right)$.

Bishop's PRML Fig 3.7

Example in 1-Dimension: 1 Observation

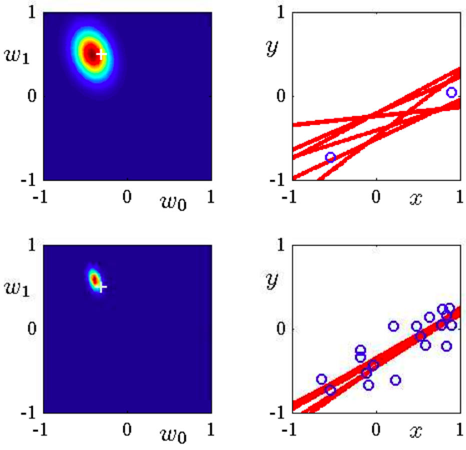




- On left: posterior distribution; white cross indicates true parameters
- On right:
  - blue circle indicates the training observation
  - red lines, $y(x) = \mathbb{E}[y \mid x, w] = w_0 + w_1 x$, for randomly chosen $w \sim p(w|\mathcal{D})$ (posterior)

Bishop's PRML Fig 3.7

Example in 1-Dimension: 2 and 20 Observations



Bishop's PRML Fig 3.7

Gaussian Regression: Closed form

Closed Form for Posterior

- Model:

$$\begin{aligned} w &\sim \mathcal{N}(0, \Sigma_0) \\ y_i \mid x, w &\text{ i.i.d. } \mathcal{N}(w^T x_i, \sigma^2) \end{aligned}$$

- Design matrix X Response column vector y
- **Posterior distribution is a Gaussian distribution:**

$$\begin{aligned} w \mid \mathcal{D} &\sim \mathcal{N}(\mu_P, \Sigma_P) \\ \mu_\mathbf{P} &= \left(X^T X + \sigma^2 \Sigma_0^{-1}\right)^{-1} X^T y \\ \Sigma_\mathbf{P} &= \left(\sigma^{-2} X^T X + \Sigma_0^{-1}\right)^{-1} \end{aligned}$$

- **Posterior Variance** Σ_P gives us a natural **uncertainty measure.**

Closed Form for Posterior

- **Posterior distribution is a Gaussian distribution:**

$$
\begin{aligned}
w \mid \mathcal{D} &\sim \mathcal{N}(\mu_P, \Sigma_P) \\
\mu_{\mathbf{P}} &= \left(X^T X + \sigma^2 \Sigma_0^{-1}\right)^{-1} X^T y \\
\Sigma_{\mathbf{P}} &= \left(\sigma^{-2} X^T X + \Sigma_0^{-1}\right)^{-1}
\end{aligned}
$$

- If we want point estimates of w, **MAP estimator** and the **posterior mean** are given by

$$
\hat{w} = \mu_P = \left(X^T X + \sigma^2 \Sigma_0^{-1}\right)^{-1} X^T y
$$

- For the prior variance $\Sigma_0 = \frac{\sigma^2}{\lambda} I$, we get

$$
\hat{w} = \mu_P = \left(X^T X + \lambda I\right)^{-1} X^T y,
$$

which is of course the ridge regression solution.