Some relevant formulas

Pages 1–3 contain some expressions that may or may not be useful for solving the exam problems. *This is not a complete list of formulas used in the course*, but some of the problems may require knowledge about certain expressions not listed here. Furthermore, the formulas listed below *are not self-explanatory*, meaning that you need to be familiar with the expressions to be able to interpret them. They are possibly a support for solving the problems, but *not* a comprehensive summary of the course.

The Gaussian distribution: The probability density function of the p-dimensional Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ is

$$\mathcal{N}(x \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{p/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(x-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(x-\boldsymbol{\mu})\right), \qquad x \in \mathbb{R}^p.$$

Sum of identically distributed variables: For identically distributed random variables $\{z_i\}_{i=1}^n$ with mean $\boldsymbol{\mu}$, variance σ^2 and average correlation between distinct variables ρ, it holds that $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n z_i\right] = \boldsymbol{\mu}$ and $\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n z_i\right) = \frac{1-\rho}{n}\sigma^2 + \rho\sigma^2$.

Linear regression and regularization:

- The least-squares estimate of $\boldsymbol{\beta}$ in the linear regression model

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \epsilon$$

  is given by the solution $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ to the normal equations $\mathbf{X}^{\mathsf{T}}\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} = \mathbf{X}^{\mathsf{T}}\mathbf{y}$, where

$$\mathbf{X} = \begin{bmatrix} 1 & -\mathbf{x}_1^{\mathsf{T}}- \\ 1 & -\mathbf{x}_2^{\mathsf{T}}- \\ \vdots & \vdots \\ 1 & -\mathbf{x}_n^{\mathsf{T}}- \end{bmatrix} \text{ and } \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \text{ from the training data } \mathcal{T} = \{\mathbf{x}_i, y_i\}_{i=1}^n$$

- Ridge regression uses the regularization term $\gamma\|\boldsymbol{\beta}\|_2^2 = \gamma\sum_{j=0}^p \beta_j^2$.
  The ridge regression estimate is $\widehat{\boldsymbol{\beta}}_{\mathrm{RR}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \gamma\mathbf{I})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$.
- LASSO uses the regularization term $\gamma\|\boldsymbol{\beta}\|_1 = \gamma\sum_{j=0}^p |\beta_j|$.

Maximum likelihood: The maximum likelihood estimate is given by

$$\widehat{\boldsymbol{\beta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\beta}} \log \ell(\boldsymbol{\beta})$$

where $\log \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log p(y_i \mid \mathbf{x}_i; \boldsymbol{\beta})$ is the log-likelihood function (the last equality holds when the n training data points are modeled to be independent).

Logistic regression: The logistic regression combines linear regression with the logistic function to model the class probability

$$p(y = 1 \mid \mathbf{x}) = \frac{e^{\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}}}{1 + e^{\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}}}.$$

For multi-class logistic regression we use the *softmax* function and models

$$p(y = k \mid \mathbf{x}_i) = \frac{e^{\boldsymbol{\beta}_k^{\mathsf{T}}\mathbf{x}_i}}{\sum_{l=1}^{K} e^{\boldsymbol{\beta}_l^{\mathsf{T}}\mathbf{x}_i}}.$$

Discriminant Analysis: The linear discriminant analysis (LDA) classifier models $p(y \mid \mathbf{x})$ using Bayes' theorem and the following assumptions

$$p(y = k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid k)p(k)}{\sum_{j=1}^{K} p(\mathbf{x} \mid j)p(j)} = \frac{\mathcal{N}\left(\mathbf{x} \mid \widehat{\boldsymbol{\mu}}_k, \widehat{\Sigma}\right)\widehat{\pi}_k}{\sum_{j=1}^{K} \mathcal{N}\left(\mathbf{x} \mid \widehat{\boldsymbol{\mu}}_j, \widehat{\Sigma}\right)\widehat{\pi}_j},$$

where

$$\begin{aligned}
\widehat{\pi}_k &= n_k/n \text{ for } k = 1, \ldots, K \\
\widehat{\boldsymbol{\mu}}_k &= \frac{1}{n_k} \sum_{i:y_i=k} \mathbf{x}_i \text{ for } k = 1, \ldots, K \\
\widehat{\Sigma} &= \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)^{\mathsf{T}}.
\end{aligned}$$

For quadratic discriminant analysis (QDA), the assumption is

$$p(y = k \mid \mathbf{x}) = \frac{\mathcal{N}\left(\mathbf{x} \mid \widehat{\boldsymbol{\mu}}_k, \widehat{\Sigma}_k\right)\widehat{\pi}_k}{\sum_{j=1}^{K} \mathcal{N}\left(\mathbf{x} \mid \widehat{\boldsymbol{\mu}}_j, \widehat{\Sigma}_j\right)\widehat{\pi}_j},$$

where $\widehat{\boldsymbol{\mu}}_k$ and $\widehat{\pi}_k$ are as for LDA, and

$$\widehat{\Sigma}_k = \frac{1}{n-1} \sum_{i:y_i=k} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)^\mathsf{T}.$$

Classification trees: The cost function for tree splitting is $\sum_{m=1}^{|T|} n_m Q_m$ where T is the tree, $|T|$ the number of terminal nodes, n_m the number of training data points falling in node m, and Q_m the impurity of node m. Three common impurity measures for splitting classification trees are:

$$\begin{aligned}
\text{Misclassification error:} \qquad & Q_m = 1 - \max_k \widehat{\pi}_{mk} \\
\text{Gini index:} \qquad & Q_m = \sum_{k=1}^{K} \widehat{\pi}_{mk}(1 - \widehat{\pi}_{mk}) \\
\text{Entropy/deviance:} \qquad & Q_m = -\sum_{k=1}^{K} \widehat{\pi}_{mk} \log \widehat{\pi}_{mk}
\end{aligned}$$

where $\widehat{\pi}_{mk} = \frac{1}{n_m} \sum_{i:\mathbf{x}_i \in R_m} \mathbb{I}(y_i = k)$

Loss functions for classification: For a binary classifier expressed as $\widehat{y}(\mathbf{x}) = \text{sign}\{C(\mathbf{x})\}$, for some real-valued function $C(\mathbf{x})$, the margin is defined as $y \cdot C(\mathbf{x})$ (note the convention $y \in \{-1, 1\}$ here). A few common loss functions expressed in terms of the margin, $L(y, C(\mathbf{x}))$ are,

$$\begin{aligned}
\text{Exponential loss:} \qquad & L(y, c) = \exp(-yc). \\
\text{Hinge loss:} \qquad & L(y, c) = \begin{cases} 1 - yc & \text{for } yc < 1, \\ 0 & \text{otherwise.} \end{cases} \\
\text{Binomial deviance:} \qquad & L(y, c) = \log(1 + \exp(-yc)). \\
\text{Huber-like loss:} \qquad & L(y, c) = \begin{cases} -yc & \text{for } yc < -1, \\ \frac{1}{4}(1 - yc)^2 & \text{for } -1 \le yc \le 0, \\ 0 & \text{otherwise.} \end{cases} \\
\text{Misclassification loss:} \qquad & L(y, c) = \begin{cases} 1 & \text{for } yc < 0, \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$